Ngoc Thanh Nguyen
Chong-Gun Kim
Adam Janiak (Eds.)

LNAI 6592

Intelligent Information and Database Systems

Third International Conference, ACIIDS 2011
Daegu, Korea, April 2011
Proceedings, Part II

2 Part II

Springer

Lecture Notes in Artificial Intelligence 6592

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Ngoc Thanh Nguyen Chong-Gun Kim
Adam Janiak (Eds.)

Intelligent Information and Database Systems

Third International Conference, ACIIDS 2011
Daegu, Korea, April 20-22, 2011
Proceedings, Part II

Springer

Volume Editors

Ngoc Thanh Nguyen
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

Chong-Gun Kim
Yeungnam University
Department of Computer Engineering
Dae-Dong, 712-749 Gyeungsan, Korea
E-mail: cgkim@yu.ac.kr

Adam Janiak
Wrocław University of Technology
Institute of Informatics, Automation and Robotics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail: adam.janiak@pwr.wroc.pl

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-20041-0 e-ISBN 978-3-642-20042-7
DOI 10.1007/978-3-642-20042-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011923233

CR Subject Classification (1998): I.2, H.3, H.2.8, H.4-5, F.1, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

ACIIDS 2011 was the third event of the series of international scientific conferences for research and applications in the field of intelligent information and database systems. The aim of ACIIDS 2011 was to provide an international forum for scientific research in the technologies and applications of intelligent information, database systems and their applications. ACIIDS 2011 was co-organized by Yeungnam University (Korea) and Wroclaw University of Technology (Poland) and took place in Deagu (Korea) during April 20–22, 2011. The first two events, ACIIDS 2009 and ACIIDS 2010, took place in Dong Hoi city and Hue city in Vietnam, respectively.

We received more than 310 papers from 27 countries over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 110 papers with the highest quality were selected for oral presentation and publication in the two volumes of ACIIDS 2011 proceedings.

The papers included in the proceedings cover the following topics: intelligent database systems, data warehouses and data mining, natural language processing and computational linguistics, Semantic Web, social networks and recommendation systems, collaborative systems and applications, e-bussiness and e-commerce systems, e-learning systems, information modeling and requirements engineering, information retrieval systems, intelligent agents and multi-agent systems, intelligent information systems, intelligent Internet systems, intelligent optimization techniques, object-relational DBMS, ontologies and knowledge sharing, semi-structured and XML database systems, unified modeling language and unified processes, Web services and Semantic Web, computer networks and communication systems.

Accepted and presented papers highlight the new trends and challenges of intelligent information and database systems. The presenters showed how new research could lead to new and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Honorary Chairs, Tadeusz Więckowski (Rector of Wroclaw University of Technology, Poland), Makoto Nagao (President of National Diet Library, Japan), and Key-Sun Choi (KAIST, Korea) for their support.

Our special thanks go to the Program Co-chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process which helped us to guarantee the highest quality of the selected papers for the conference. We cordially thank the organizers and chairs of special sessions, who essentially contribute to the success of the conference.

We also would like to express our thanks to the keynote speakers (Hamido Fujita, Halina Kwaśnicka, Yong-Woo Lee and Eugene Santos Jr.) for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, Yeungnam University (Korea), Wroclaw University of Technology (Poland) and University of Information Technology (Vietnam). Our special thanks are due also to Springer for publishing the proceedings, and the other sponsors for their kind support.

We wish to thank the members of the Organizing Committee for their very substantial work, especially those who played essential roles: Jason J. Jung, Radosław Katarzyniak (Organizing Chairs) and the members of the Local Organizing Committee for their excellent work.

We cordially thank all the authors for their valuable contributions and other participants of this conference. The conference would not have been possible without them.

Thanks are also due to many experts who contributed to making the event a success.

April 2011

Ngoc Thanh Nguyen
Chong-Gun Kim
Adam Janiak

Conference Organization

Honorary Chairs

Tadeusz Więckowski	Rector of Wroclaw University of Technology, Poland
Key-Sun Choi	KAIST, Korea
Makoto Nagao	President of National Diet Library, Japan

General Co-chairs

Chong-Gun Kim	Yeungnam University, Korea
Adam Janiak	Wroclaw University of Technology, Poland

Program Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Organizing Chairs

Jason J. Jung	Yeungnam University, Korea
Radosław Katarzyniak	Wroclaw University of Technology, Poland

Program Co-chairs

Dosam Hwang	Yeungnam University, Korea
Zbigniew Huzar	Wroclaw University of Technology, Poland
Pankoo Kim	Chosun University, Korea
Edward Szczerbicki	University of Newcastle, Australia
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Kiem Hoang	University of Information Technology, Vietnam

Special Session Chair

Bogdan Trawiński	Wroclaw University of Technology, Poland

Organizing Committee

Marcin Maleszka	Wroclaw University of Technology, Poland
Bernadetta Mianowska	Wroclaw University of Technology, Poland

Xuan Hau Pham	Yeungnam University, Korea
Adrianna Kozierkiewicz-Hetmańska	Wroclaw University of Technology, Poland
Anna Kozłowska	Wroclaw University of Technology, Poland
Wojciech Lorkiewicz	Wroclaw University of Technology, Poland
Hai Bang Truong	University of Information Technology, Vietnam
Mi-Nyer Jeon	Yeungnam University, Korea

Steering Committee

Ngoc Thanh Nguyen - Chair	Wroclaw University of Technology, Poland
Longbing Cao	University of Technology Sydney, Australia
Adam Grzech	Wroclaw University of Technology, Poland
Tu Bao Ho	Japan Advanced Institute of Science and Technology, Japan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, Korea
Jason J. Jung	Yeungnam University, Korea
Hoai An Le-Thi	Paul Verlaine University - Metz, France
Antoni Ligęza	AGH University of Science and Technology, Poland
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland

Keynote Speakers

Hamido Fujita	Iwate Prefectural University, Japan
Halina Kwaśnicka	Wroclaw University of Technology, Poland
Yong-Woo Lee	University of Seoul, Korea
Eugene Santos Jr.	Thayer School of Engineering at Dartmouth College, USA

Special Sessions Organizers

1. *Multiple Model Approach to Machine Learning (MMAML 2011)*

Oscar Cordón	European Centre for Soft Computing, Spain
Przemysław Kazienko	Wroclaw University of Technology, Poland
Bogdan Trawiński	Wroclaw University of Technology, Poland

2. *International Workshop on Intelligent Management and e-Business (IMeB 2011)*

Chulmo Koo	Chosun University, Korea
Jason J. Jung	Yeungnam University, Korea

3. *Intelligent Cloud Computing and Security (ICCS 2011)*

Hyung Jong Kim	Seoul Women's University, Korea
Young Shin Han	SungKyunKwan University, Korea

4. *Modeling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)*

Le Thi Hoai An	Paul Verlaine University – Metz, France
Pham Dinh Tao	National Institute for Applied Science-Rouen, France

5. *User Adaptive Systems for Mobile Wireless Systems (UAS 2011)*

Ondrej Krejcar	VŠB-Technical University of Ostrava, Czech Republic
Peter Brida	University of Žilina, Slovakia

6. *International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS)*

Kun Chang Lee	Sungkyunkwan University, Korea
Oh Byung Kwon	Kyunghee University, Korea
Jae Kyeong Kim	Kyunghee University, Korea

International Program Committee

El-Houssaine Aghezzaf	Ghent University, Belgium
Le Thi Hoai An	Paul Verlaine University - Metz, France
Costin Badica	University of Craiova, Romania
Maria Bielikova	Slovak University of Technology, Slovakia
Nguyen Thanh Binh	Hue University, Vietnam
Lydie Boudjeloud-Assala	Paul Verlaine University - Metz, France
Stephane Bressane	School of Computing, Singapore
Grażyna Brzykcy	Poznań University of Technology, Poland
The Duy Bui	Vietnam National University, Vietnam
Longbing Cao	University of Technology, Sydney, Australia
Frantisek Capkovic	Slovak Academy of Sciences, Slovakia
Oscar Castillo	Tijuana Institute of Technology, Mexico
Wooi Ping Cheah	Multimedia University, Malaysia
Jr-Shian Chen	Hungkuang University, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Tzu-Fu Chiu	Aletheia University, Taiwan
Kyung-Yong Chung	Sangji University, Korea
Alfredo Cuzzocrea	University of Calabria, Italy

Phan Cong-Vinh	London South Bank University, UK
Ireneusz Czrnowski	Gdynia Maritime University, Poland
Tran Khanh Dang	National University of Ho Chi Minh City, Vietnam
Paul Davidsson	Blekinge Institute of Technology, Sweden
Victor Felea	Alexandru Ioan Cuza University of Iasi, Romania
Jesus Alcala Fernandez	Universidad de Granada, Spain
Pawel Forczmanski	Zachodniopomorski Technical University, Poland
Dariusz Frejlichowski	Zachodniopomorski Technical University, Poland
Patrick Gallinar	UPMC, France
Irene Garrigós	University of Alicante, Spain
Hoang Huu Hanh	Vienna University of Technology, Austria
Jin-Kao Hao	University of Angers, France
Heikki Helin	University of Helsinki, Finland
Kiem Hoang	University of Information Technology, Vietnam
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Chenn-Jung Huang	National Dong Hwa University, Taiwan
Gordan Jezic	University of Zagreb, Croatia
Robert Kapłon	Wroclaw University of Technology, Poland
Shuaib Karim	Quaid-i-Azam University, Pakistan
Radosław Katarzyniak	Wroclaw University of Technology, Poland
Przemysław Kazienko	Wroclaw University of Technology, Poland
Cheonshik Kim	Sejong University, Korea
Joanna Kolodziej	University of Bielsko-Biala, Poland
Romuald Kotowski	Polish-Japanese Institute of Information Technology, Poland
Ondrej Krejcar	Technical University of Ostrava, Czech Republic
Dariusz Król	Wroclaw University of Technology, Poland
Tomasz Kubik	Wroclaw University of Technology, Poland
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Halina Kwaśnicka	Wroclaw University of Technology, Poland
Raymond Lau	City University of Hong Kong, China
Eunser Lee	Andong National University, Korea
Zne-Jung Lee	National Taiwan University of Science and Technology, Taiwan
Chunshien Li	National Central University, Taiwan
Jose-Norberto Mazón López	University of Alicante, Spain
Marie Luong	Université Paris 13 Nord, France
Urszula Markowska-Kaczmar	Wroclaw University of Technology, Poland
Tadeusz Morzy	Poznań University of Technology, Poland
Kazumi Nakamatsu	University of Hyogo, Japan

Phi Khu Nguyen	University of Information Technology, Vietnam
Grzegorz Nalepa	AGH University of Science and Technology, Poland
Vincent Nguyen	The University of New South Wales, Australia
Toyoaki Nishida	Kyoto University, Japan
Cezary Orłowski	Gdansk University of Technology, Poland
Chung-Ming Ou	Kainan University, Taiwan
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Marcin Paprzycki	Systems Research Institute of the Polish Academy of Sciences, Poland
Do Phuc	University of Information Technology, Vietnam
Bhanu Prasad	Florida Agricultural and Mechanical University, USA
Witold Rekuć	Wroclaw University of Technology, Poland
Ibrahima Sakho	Paul Verlaine University - Metz, France
An-Zen Shih	Jinwen University of Science and Technology, Taiwan
Janusz Sobecki	Wroclaw University of Technology, Poland
Serge Stinckwich	Université de Caen Basse Normandie, France
Pham Dinh Tao	INSA of Rouen, France
Wojciech Thomas	Wroclaw University of Technology, Poland
Bogdan Trawiński	Wroclaw University of Technology, Poland
Hoang Hon Trinh	Ho Chi Minh City University of Technology, Vietnam
Hong-Linh Truong	Vienna University of Technology, Austria
K. Vidyasankar	Memorial University, Canada
Jia-Wen Wang	Nanhua University, Taiwan
Michal Wozniak	Wroclaw University of Technology, Poland
Xin-She Yang	National Physical Laboratory, UK
Wang Yongli	North China Electric Power University, China
Zhongwei Zhang	University of Southern Queensland, Australia

Program Committees of Special Sessions

Special Session on Multiple Model Approach to Machine Learning (MMAML 2011)

Hussein A. Abbass	University of New South Wales, Australia
Ajith Abraham	Norwegian University of Science and Technology, Norway
Jesús Alcalá-Fdez	University of Granada, Spain
Ethem Alpaydin	Bogaziçi University, Turkey
Oscar Castillo	Tijuana Institute of Technology, Mexico
Rung-Ching Chen	Chaoyang University of Technology, Taiwan

Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Emilio Corchado	University of Burgos, Spain
Oscar Cordón	European Centre for Soft Computing, Spain
José Alfredo F. Costa	Federal University (UFRN), Brazil
Mustafa Mat Deris	University Tun Hussein Onn Malaysia, Malaysia
Patrick Gallinari	Pierre et Marie Curie University, France
Lawrence O. Hall	University of South Florida, USA
Francisco Herrera	University of Granada, Spain
Frank Hoffmann	TU Dortmund, Germany
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Hisao Ishibuchi	Osaka Prefecture University, Japan
Yaochu Jin	University of Surrey, UK
Przemysław Kazienko	Wrocław University of Technology, Poland
Yong Seog Kim	Utah State University, USA
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Kin Keung Lai	City University of Hong Kong, Hong Kong
Mark Last	Ben-Gurion University of the Negev, Israel
Kun Chang Lee	Sungkyunkwan University, Korea
Chunshien Li	National Central University, Taiwan
Heitor S. Lopes	Federal University of Technology Paraná, Brazil
Edwin Lughofer	Johannes Kepler University Linz, Austria
Witold Pedrycz	University of Alberta, Canada
Ke Tang	University of Science and Technology of China, China
Bogdan Trawiński	Wrocław University of Technology, Poland
Olgierd Unold	Wrocław University of Technology, Poland
Michał Wozniak	Wrocław University of Technology, Poland
Zhongwei Zhang	University of Southern Queensland, Australia
Zhi-Hua Zhou	Nanjing University, China

The Third International Workshop on Intelligent Management and e-Business (IMeB 2011)

Chang E. Koh	University of North Texas, USA
Bomil Suh	Sookmyung Women's University, Korea
Hee-Woong Kim	Yonsei University, Korea
Vijay Sugumaran	Oakland University, USA
Angsana Techatassanasoontorn	Pennsylvania State University, USA
Yong Kim	Utah State University, USA
Shu-Chun Ho	National Kaohsiung Normal University, Taiwan
Jae-Nam Lee	Korea University, Korea
Namho Chung	Kyung Hee University, Korea
Joon Koh	Chonnam National University, Korea

Sang-Jun Lee	Chonnam National University, Korea
Kichan Nam	Sogang University, Korea
Bo Choi	Vanderbilt University, USA
GeeWoo Bock	SungKyungwan University, Korea
Dahui Li	University of Minnesota Duluth, USA
Jahyun Goo	Florida Atlantic University, USA
Shane Tombline	Marshall University, USA
Dale Shao	Marshall University, USA
Jae-Kwon Bae	Dongyang University, Korea
Hyunsoo Byun	Backsuk Art University, Korea
Jaeki Song	Texas Tech University, USA
Yongjin Kim	Sogang University, Korea

The Special Session on Intelligent Cloud Computing and Security (ICCS 2011)

Jemal. H. Abawajy	Deakin University, Australia
Ahmet Koltuksuz	Izmir Institute of Technology, Turkey
Brian King	Indiana University Purdue University Indianapolis, USA
Tatsuya Akutsu	Kyoto University, Japan
Minjeong Kim	University of North Carolina, USA
Hae Sang Song	Seowon University, Korea
Won Whoi Huh	Sungkyul University, Korea
Jong Sik Lee	Inha University, Korea
Chungman Seo	University of Arizona, USA
Il-Chul Moon	KAIST, Korea
Hyun Suk Cho	ETRI, Korea
JI Young Park	Ewha Women's University, Korea
Tae-Hoon Kim	Hannam University, Korea
Hee Suk Suh	Korea University of Technology, Korea
Lei SHU	Osaka University, Japan
Mukaddim Pathan	CSIRO, Australia
Al-Sakib Khan Pathan	International Islamic University Malaysia

Recent Advances in Modeling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)

El-Houssaine Aghezzaf	Ghent University, Belgium
Le Thi Hoai An	Paul Verlaine University of Metz, France
Lydie Boudjeloud-Assala	Paul Verlaine University of Metz, France
Brieu Conan-Guez	Paul Verlaine University of Metz, France
Alain Gely	Paul Verlaine University of Metz, France

Jin-Kao Hao	University of Angers, France
Proth Jean-Marie	INRIA-Metz, France
Francois-Xavier Jollois	University of Paris V, France
Marie Luong	University of Paris 13, France
Do Thanh Nghi	Enst Brest, France
Vincent Nguyen	The University of New South Wales, Australia
Ibrahima Sakho	Paul Verlaine University of Metz, France
Daniel Singer	Paul Verlaine University of Metz, France
Pham Dinh Tao	Insa of Rouen, France
Bogdan Trawiński	Wrocław University of Technology, Poland

Special Session on User Adaptive Systems for Mobile Wireless Systems (UAS 2011)

Tipu Arvind Ramrekha	Kingston University, London, UK
Peter Brida	University of Zilina, Slovakia
Wei-Chen Cheng	Academia Sinica, Taiwan, Republic of China
Theofilos Chrysikos	University of Patras, Greece
Michael Feld	German Research Center for Artificial Intelligence (DFKI), Germany
Robert Frischer	VSB Technical University of Ostrava, Czech Republic
Jiri Horak	VSB Technical University of Ostrava, Czech Republic
Vladimir Kasik	VSB Technical University of Ostrava, Czech Republic
Aleksander Kostuch	Politechnika Gdanska / Sprint Sp. z o.o., Poland
Stavros Kotsopoulos	University of Patras, Greece
Jiri Kotzian	VSB Technical University of Ostrava, Czech Republic
Ondrej Krejcar	VSB Technical University of Ostrava, Czech Republic
Luca Longo	Trinity College Dublin, Republic of Ireland
Zdenek Machacek	VSB Technical University of Ostrava, Czech Republic
Juraj Machaj	University of Zilina, Slovakia
Norbert Majer	Research Institute of Posts and Telecommunications, Slovakia
Rainer Mautz	Swiss Federal Institute of Technology, Zurich, Switzerland
Marek Penhaker	VSB Technical University of Ostrava, Czech Republic
Stefan Pollak	University of Zilina, Slovakia
Bogdan Trawiński	Wrocław University of Technology, Poland
Jan Vyjidak	Cardiff University, UK
Vladimir Wieser	University of Zilina, Slovakia

International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS 2011)

Hyunchul Ahn	Kookmin University, Korea
Michael Beigl	Karlsruhe Institute of Technology, Germany
Seong Wook Chae	NIA, Korea
Namyong Cho	Samsung SDS Co., Korea
Do Young Choi	LG CNS, Korea
Hyng Seung Choo	Sungkyunkwan University, Korea
Namho Chung	Kyung Hee University, Korea
Avelino J. Gonzalez	University of Central Florida, USA
Yousub Hwang	University of Seoul, Korea
Hyea Kyeong Kim	Kyung Hee University, Korea
Jae Kyeong Kim	Kyung Hee University, Korea
Namgyu Kim	Kookmin University, Korea
Hye-Kyeong Ko	Korea Advanced Institute of Science and Technology, Korea
Ohbyung Kwon	Kyung Hee University, Korea
Dongwon Lee	Korea University, Korea
Hyun Jeong Lee	Korea University, Korea
Kun Chang Lee	Sungkyunkwan University, Korea
Sang Ho Lee	Sun Moon University, Korea
Sangjae Lee	Sejong University, Korea
Bong Won Park	Sungkyunkwan University, Korea
Hedda R. Schmidtke	Karlsruhe Institute of Technology (KIT), Germany
Young Wook Seo	NIPA, Republic of Korea
Stephan Sigg	Institute of Operating Systems and Computer Networks, Germany
Bogdan Trawiński	Wrocław University of Technology, Poland

Additional Reviewers

Trong Hai Duong	Inha University, Korea
Bernadetta Mianowska	Wroclaw University of Technology, Poland
Michał Sajkowski	Poznan University of Technology, Poland
Robert Susmaga	Poznan University of Technology, Poland

Table of Contents – Part II

Intelligent Optimization Techniques

Rough Set Based and Fuzzy Set Based Systems

Intelligent Information Retrieval

Computer Vision Techniques

Multiple Model Approach to Machine Learning (MMAML 2011)

Intelligent Cloud Computing and Security (ICCS 2011)

Modelling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)

User Adaptive Systems for Mobile Wireless Systems (UAS 2011)

International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS)

Table of Contents – Part I

Keynote Speeches

Intelligent Database Systems

Data Warehouses and Data Mining

Natural Language Processing and Computational Linguistics

Semantic Web, Social Networks and Recommendation Systems

Technologies for Intelligent Information Systems

Collaborative Systems and Applications

E-Business and e-Commerce Systems

Evolutionary Algorithms for Base Station Placement in Mobile Networks

Piotr Regula, Iwona Pozniak-Koszalka, Leszek Koszalka, and Andrzej Kasprzak

Dept. of Systems and Computer Networks, Wroclaw University
of Technology, 50-370 Wroclaw, Poland
leszek.koszalka@pwr.wroc.pl

Abstract. Base station (BS) placement has a significant impact on the performance of mobile networks. Currently several algorithms are in use, differing in their complexity, task allocation time, and memory usage. In this paper, base station placement is automatically determined using two algorithms: a genetic, one and a hybrid algorithm, which combines both genetic and tabu search metaheuristics. Both designed algorithms are described, and compared with each other. For the positioning of the base station, both the power of the BS, and the size (location) of the subscriber group (SG) were considered factors. Additionally, the question of how the algorithms' parameters influence the solution was investigated. In order to conduct the investigation, a special application was created that runs both algorithms. Simulation tests prove that the hybrid algorithm outperforms the genetic algorithm in most cases. The hybrid algorithm delivers near-optimal solutions.

Keywords: mobile network, base station placement, optimization, evolutionary algorithm, experimental system.

1 Introduction

The question of base stations placement is one of the most important in the area of designing mobile networks. There exist algorithms for base station placement in wireless networks based on various approaches described in [1], [2], [3], and [4]. Some of them are based on the ideas of evolutionary algorithms [5], [6], and [7]. In the paper, we focus on the design and implementation of our own hybrid evolutionary algorithm. The evaluation of this algorithm and the comparison with an implementation of a genetic algorithm is made using our own experimental system described in the paper.

The rest of the paper is organized as follows: Problem statement is in Section 2, with definition of input, and output parameters of the problem, cost function and the criterion for comparing the algorithms. In Section 3, the implemented algorithms for BS placement are described, including genetic algorithm (GA), tabu search (TS), and hybrid (HA). Application overview in Section 4 presents our created experimental system, which can simulate the process of positioning base stations, with the possibility of specifying the inputs, and displaying complete output statistics. Some results of research with introduction to the statistical comparison of the algorithms are in Section 5. Conclusions and perspectives appear in Section 6.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 1–10, 2011.

2 Problem Statement

The main subject of this paper is the positioning of a BS in a three-dimensional service area. The service area is divided into a large number of pixels. Each pixel is a potential position for new BS. The positioning of a new BS takes into consideration factors such as: the power and position of other existing BS, as well as the sizes (total number in the area and quantities within groups) and locations of the subscriber groups (SG) lying in the service area. The problem parameters are listed in Table 1.

Table 1. Parameters in Base Station Placement problem

Symbol	Parameter
k	Number of existing BS
$\mathbf{P_i}$	Power of Existing BSi
$\mathbf{P_n}$	Power of the New BS
$\mathbf{(q_i,j_i,l_i)}$	Position of existing BSi
$\mathbf{(w_i,z_i,n_i)}$	Position of existing SGi
s	Number of SG
$\mathbf{N_i}$	Quantity of members in SGi
m	Number of new BS
$\mathbf{(x_i,y_i,h_i)}$	Coordinates of new base stations (position of new BS)
r_i	Distance between new, and existing BS

The proposed algorithms are verified by a cost function. An ideal positioning algorithm would result in selecting appropriate new-allocated BSs location (configuration). It means a new BS should be placed close to its SG, and as far away as possible from existing BS. In the paper, an introduced cost function is based on this interpretation and is expressed as a sum of p coefficients (1).

$$f(x_1, y_1, h_1, \ldots, x_m, y_m, h_m) = p_1 + \ldots + p_m \tag{1}$$

Coefficient p is calculated for each new BS, regardless of the position of any other BS

$$p_i = \alpha(r_1 P_1 + r_2 P_2 + \cdots r_{k+m-1} P_{k+m-1}) - \\ -\beta(r'_1 N_1 + r'_2 N_2 + \cdots r'_F N_F) \tag{2}$$

where:

$$r_i = \sqrt{(x_i - q_i)^2 + (y_i - j_i)^2 + (h_i - l_i)^2} \tag{3}$$

$$r'_i = \sqrt{(x_i - w_i)^2 + (y_i - z_i)^2 + (h_i - n_i)^2} \tag{4}$$

The greater the value of the cost function, the better is the BS configuration. Fig. 1 illustrates a graphic representation of a typical service area. Existing BSs are marked in white color, whereas new BSs in green. Some exemplary parameters (power and position) are written below for every BS.

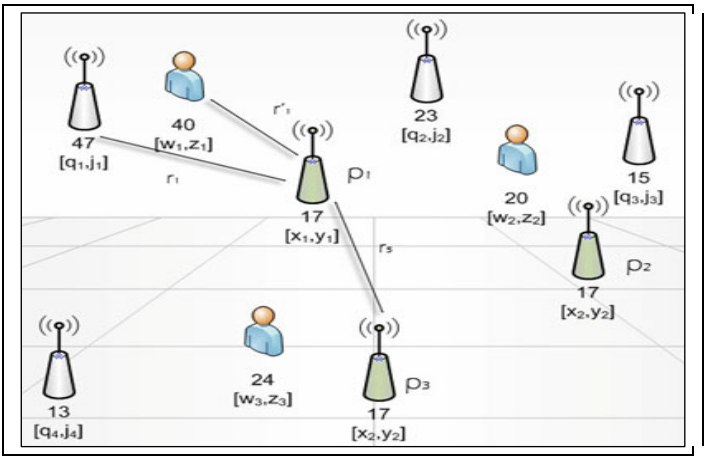


Fig. 1. Service area – an example

3 Algorithms

The objective of this research is to optimize BS placement. The problem has a very large solution space, so the only practical way to find good solutions to search for them using metaheuristic algorithms. BS placement problems cannot be solved in polynomial time. So, in this paper, BS placement is achieved using genetic, and tabu search algorithms. These algorithms are useful for solving NP-hard problems. Both algorithms are described in details below.

Genetic algorithm (GA). Genetic algorithms are evolutionary optimization approaches (e.g. [8] and [9]) that are an alternative to traditional optimization algorithms. GAs find solutions by exploiting dynamic principles governing natural selection and genetic processes, e.g. recombination and mutation. Implementation of a GA is highly dependent on the type of problem considered. Therefore, for the BS location problem the following assumptions have been made:

- chromosomes (creatures) correspond to the positions of all the new BS.
- the fitness (adaptation) function (required to evaluate the solution domain) corresponds to the cost function described in the previous Section.

Initially, a GA randomly generates a population of random creatures. Each creature represents a potential solution to the positioning problem. The implemented algorithm consists of the following steps:

Step 1. Evaluate creature, and assign an adaptation score to it. Creatures with higher scores (evaluated by the adaptation function) have a greater probability of being selected.
Step 2. Select a new population from the current population, by the roulette wheel method.
Step 3. Crossover selected creatures. Below is shown the procedure of crossover operation used in the implementation of GA algorithm.

Dad	(2,3,2),	(3,4,3),	(1,3,1),	(7,8,6)
Mom	(3,4,2),	(5,7,3),	(5,3,1),	(8,2,6)

- randomize MinPoz, and MaxPoz,

	MinPoz		MaxPoz
0	1	2	3
(2,3,2), |	**(3,4,3),**	**(1,3,1),**	**(7,8,6),**
(3,4,2), |	**(5,7,3),**	**(5,3,1),**	**(8,2,6)**

- crossover appropriate BS.

(2,3,2), |	**(5,7,3),**	**(5,3,1),**	**(8,2,6)**
(3,4,2), |	**(3,4,3),**	**(1,3,1),**	**(7,8,6)**

Step 4. Select (with probability P_{mut}) creatures for a mutation procedure from the current population, by the roulette wheel method. Randomize BS coordinates of selected creatures.
Step 5. Repeat Step 1 to Step 4 for N times (the user defines N).

The parameters for the implemented genetic algorithm are listed in Table 2.

Table 2. Parameters in the implementation of genetic algorithm

Symbol	Parameter
S	Population size
G	Number of generations
$\mathbf{P_{cro}}$	Probability of Crossover
$\mathbf{P_{mut}}$	Probability of Mutation

Tabu Search (TS). Tabu search is an intelligent metaheuristic algorithm (e.g. [10] and [11]). Its main objective is to avoid searches that repeatedly explore the same parts of the solution space. It does this by forbidding or penalizing moves that lead to places already visited. Before describing the actual algorithm, we must clarify how some of the features of TS are implemented.

The Neighborhood of a point, in which BS is located, is defined as a sphere around that point, with radius defined by the parameter called Neighborhood Range. A *tabu list* is a list of points which were already visited. To avoid going to the same locations, the algorithm checks if a new neighboring point is not located near one of the points saved in the *tabu list* (the distance must be less than a defined variable – *tabu range*). The *tabu list* length is limited, and when its capacity is reached then new points will replace the old ones. The algorithm implemented for our specific problem is described as follows:

Step 1. Choose an initial solution from the solution space, namely a vector of selected points from a defined area.

Step 2. For each specified BS position defined by the current solution, find new, candidate positions within the BS Neighborhood.

Step 3. Check the cost function for solutions, which include the new BS positions. Find the best neighbor not listed on the *tabu list*.

Step 4. Add previously (in Step 3) checked (but not 'best neighbor') candidates to solutions to a *tabu list*, and modify the current solution by including best neighbor as one of the points in solution vector.

Step 5. Check, via the cost function, if the current solution is better than all previously-obtained 'best solutions'.

Step 6. Go back to step 1, unless the current best solution is acceptable.

Remark: To avoid congestion, the implemented TS algorithm has an additional mechanism which exceptionally (if all current neighbors are included in *tabu* list) permits the use of a point belonging to *tabu list*. Additionally, the implemented tabu search algorithm includes an accelerator which 'dramatically' increases the speed of computing the cost function for all the neighbors in Step 3.

Hybrid Algorithm (HA). Our hybrid algorithm combines both the genetic and the tabu search approaches in order to improve search performance. Since the genetic algorithm works better in the initial phase of the search, we use it to create a population and identify a good starting solution for tabu search. Once it is found, TS begins with it. These both algorithms work separately, but what really makes our algorithm a hybrid one is that the mutation process from the genetic algorithm are embedded in the tabu search diversification system enhancing its capabilities. A normal TS lacks a mechanism that radically changes the searching area (i.e. beyond the neighborhood range). Mutation processes introduce an element of randomness to this algorithm. In each iteration, TS has a small chance of mutating its current solution, so it can then start searching in a completely different part of the solution space.

To make the mutation process really efficient, in the hybrid algorithm, the probability of mutating the current solution in our modified TS phase depends on the search results themselves. If the algorithm finds solutions having similar optimization scores, the probability of mutation increases. If the algorithm consistently finds better solutions, the probability of mutation is set to its small, original value. This way mutation will not interrupt a trend of consistently finding ever better solutions, or an optimal solution. Similarly, increasing the probability of mutation prevents the algorithm from getting into the state of searching stagnation by changing the searching area drastically. Our hybrid algorithm also works much faster than the genetic one,

because it uses a speed accelerator which relies on the fact that solutions created in two consecutive iterations are very similar to each other, so our algorithm stores some more information about previous solutions gaining more speed from calculation process reduction. All parameters for hybrid algorithm are listed in Tab. 3.

Table 3. Parameters in the implementation of Hybrid Algorithm

Symbol	Parameter
I	Number of iterations
NN	Number of Neighbors
NR	Neighbourhood Rangę
TR	Tabu Range
SSGM	Starting Solution Generating Method
S	Population size
$\mathbf{P_{mut}}$	Probability of Mutation

4 Experimental System

Fig. 2 displays the logical model of the investigations, i.e. the experimentation system as input-output system designed for the purposes of this paper. It shows types of input, and output data, as well algorithm and system parameters (the notation is as described in Tab. 1, Tab. 2 and Tab. 3).

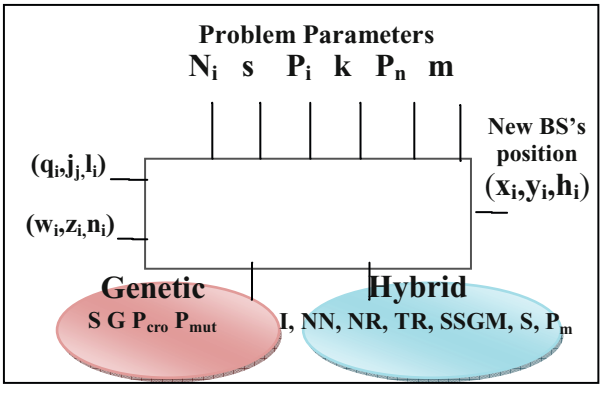


Fig. 2. Input-Output experimentation system

In order to test the considered algorithms, a special application was designed and implemented. The application simulates the performance of TS, GA, and HA as described in Section 3. The application has been created using Microsoft Visual Studio 2010 and C# language using concepts presented in [12] and [13]. Fig. 3(a) illustrates the screenshot of the application window. The generated location of BSs is presented on the left side of the window in Fig. 3(b).

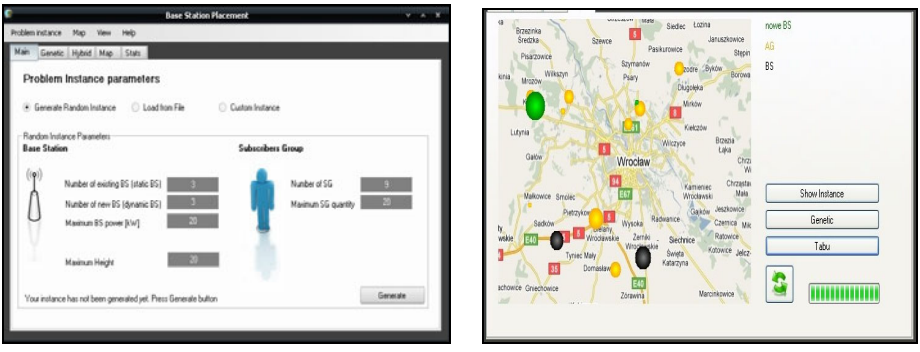
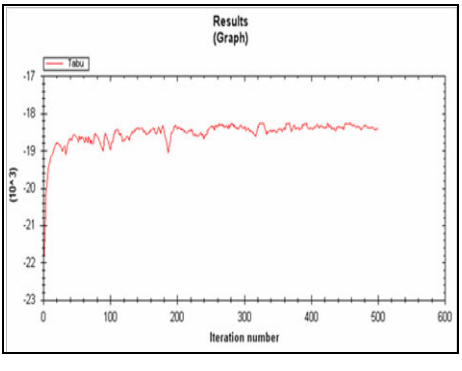


Fig. 3. Application – (a) the application window (left) and (b) the *Map* tab window (right)

The application menu provides access to four options: (i) Save or load problem instance, (ii) Load image used as a background of service area, (iii) Show graph for algorithm steps, (iv) Request help. Below the application menu there appears a special tabbed menu. The *Main* tab allows the user to select different ways of initializing the problem instance. If *Generate Random Instance* is selected, other, special parameters may be set: Number of existing and new BS, as well as maximum BS power, and Number and Quantity of SG. The *Genetic* tab allows the user to customize the parameters of the GA (e.g. Population size or Number of generations). Similarly the *Hybrid* tab allows the user to customize the parameters of the HA (e.g. Iteration number or Number of Neighbors). In the *Map* tab, the user can observe action of each algorithm.

5 Trial Test Results

Experiment #1. In the first part of the research we investigated fluctuations in the optimization score during tests. Fig. 4 illustrates the variation of optimization score (vertical axis) for HA (i.e. TS but with initialization with GA), while Fig. 5 for GA.



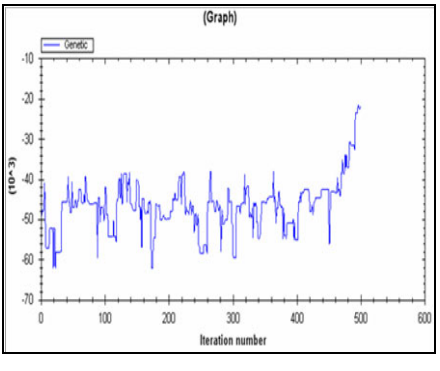


Fig. 4. Variation of optimization scores for HA **Fig. 5.** Variation of optimization score for GA

Experiment #2. The next part of the investigations was a comparison of HA and GA, depending on service area parameters: Number of existing BS (Fig. 6); Number of new BS (Fig. 7); Number of SG (Fig. 8); and Quantity of SG (Fig. 9).

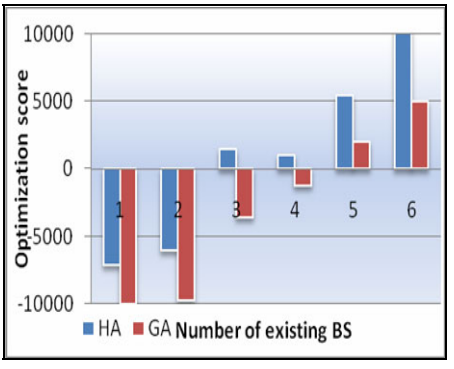


Fig. 6. Optim. score and Number of existing BS

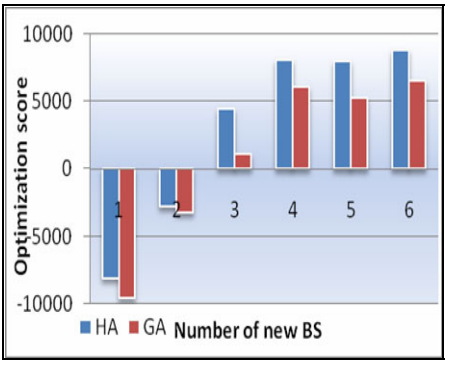


Fig. 7. Optim. score and Number of newBS

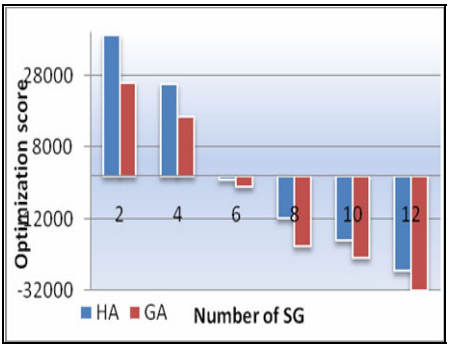


Fig. 8. Optim. score and Number of SG

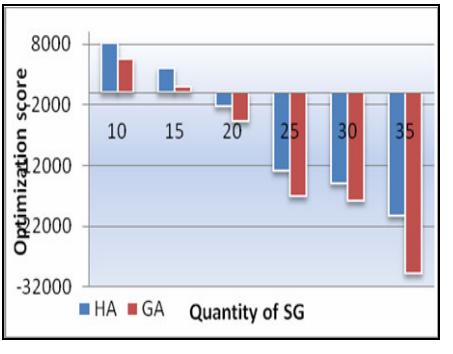


Fig. 9. Optim. score and Quantity of SG

In the figures above (where Optim. score means the value of cost function), it is apparent, that HA significantly outperforms GA. This observation was confirmed for all the tested problem instances, and for every considered set of parameter values.

Experiment #3. The investigations focus on parameters which define the neighborhood range and the range of the *tabu list*. The search was conducted for values between 20-50 pixels for Neighborhood range and values of 10-40 pixels for tabu range. The obtained results (Optim. score values) for these distinct cases are shown in Fig. 10, Fig. 11, Fig. 12, and Fig. 13. It may be observed, that for both (Neighborhood and tabu) ranges, their small values can not lead to better scores.

A low Neighborhood range (see Fig. 10) means that we check the solution space more precisely, which might give good result. However, a low range may also require more iterations to get satisfying result. It is clearly depicted by the blue

line in Fig. 11, where the graph initially attains mediocre values, but consistently finds better solutions, after a critical number of iterations.

The graph in Fig. 12 shows the influence of the *Tabu range* parameter. Setting this parameter with a high value results in algorithm behavior no different from standard random search. Moreover, a big tabu range may cause needs for using neglected tabu search mode.

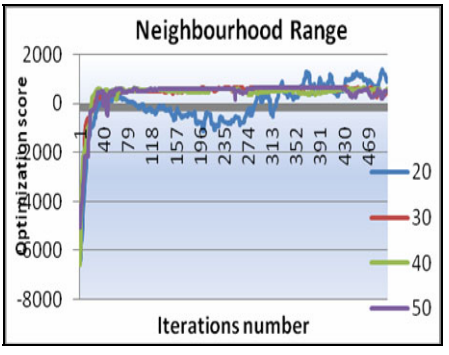


Fig. 10. Optim. score depending on Neighborhood Range

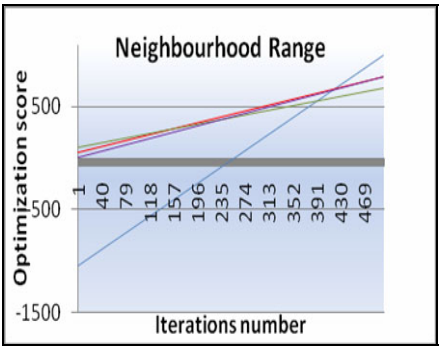


Fig. 11. Optim. score depending on Neighborhood Range (trends lines)

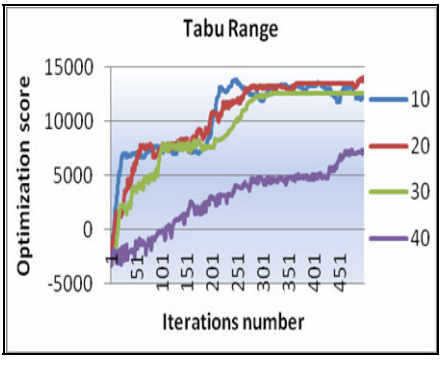


Fig. 12. Optim. score depending on Tabu Range

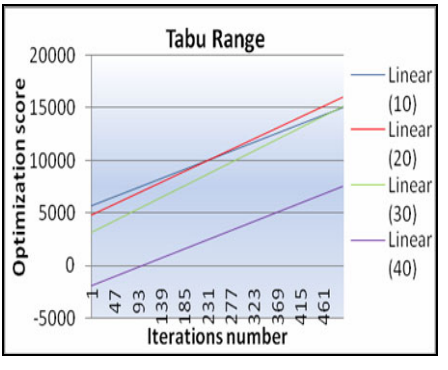


Fig. 13. Optim. score depending on Tabu Range (trends lines)

6 Conclusions

In this paper, two algorithms to base station placement in mobile network were designed and implemented, including the genetic algorithm (GA) and hybrid algorithm (HA) being a composition of elements of GA and TS. The hybrid algorithm is characterized by a smart search method, with intelligent diversification. Both algorithms have been evaluated using specially prepared experimentation system following author's ideas described in [14]. The obtained results of investigations justify the statement that HA significantly outperforms GA, in all cases.

In conclusion, this paper shows that our hybrid algorithm has truly large potential, most particularly for base station positioning.

A particularly important aspect of the designed experimentation system is the possibility to enter a lot of input data and to examine their impact on the cost function. Recently, the system is serving as a tool to aid teaching students and preparing projects in computer science and telecommunications areas in Wroclaw University of Technology.

In the further research in this area we are planning to develop testing environment by implementing ideas of multistage experiment [15] and by implementing more algorithms, including those described in [16]. Moreover, we are planning to create a hybrid algorithm developed by applying some operations from simulated annealing algorithm [9].

References

1. Ouzineb, M., Nourelfath, M., Gendreau, M.: An Efficient Heuristic for Reliability Design. Optimization Problems. Computers 37, 223–235 (2010)
2. Dorigo, M., Di Caro, G., Gambardella, L.M.: Algorithms for Discrete Optimization. University Libre de Bruxelles, Belgium (2003)
3. Rodrigues, R.C., Mateus, G.R., Loureiro, A.A.F.: Optimal Base Station Placement and Fixed Channel Assignment Applied to Wireless Local Area Network Project. In: Proc. Int. IEEE Conf. ICON, pp. 186–192 (1999)
4. Song, H.Y.: A Method of Mobile Base Station Placement for High Attitude Platform Based Network. In: Proc. of Intern. Multi-Conference on Computer Science and Information Technology, pp. 869–876 (2008)
5. Hashimuze, A., Mineno, H., Mizuno, T.: Multi-base Station Placement for Wireless Reprogramming in Sensor Networks. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5712, pp. 648–655. Springer, Heidelberg (2009)
6. Shi, Y., Hou, Y.T., Efrat, A.: Algorithm Design for Base Station Placement Problems in Sensor Networks. In: Proc. ACM Intern. Conf. Series, vol. 191 (2006)
7. Choi, Y.S., Kim, K.S.: The Displacement of Base Station in Mobile Communication with Genetic Approach. ETRI South Korea (2008)
8. Dias, A.H.F., Vasconcelos, J.A.: Multi-objective Genetic Algorithms Applied to Solve Optimization Problems. IEEE Trans. on Magn. 38, 1133–1138 (2002)
9. Tamilarasi, A., Kumar, T.A.: An Enhanced Genetic Algorithm with Simulated Annealing for Job Shop Scheduling. Int. J. of Engineering, Science and Technology 2, 141–151 (2010)
10. Glover, F., Laguna, M.: Tabu Search. Kluwer, Dordrecht (1996)
11. Gendreau, M.: An Introduction to Tabu Search. Université de Montréal (2003)
12. Hayder, H.: Object Programming with PHP 5. Helion, Gliwice (2009) (in Polish)
13. http://code.google.com/intl/pl/apis/maps/
14. Kmiecik, W., Wojcikowski, M., Koszalka, L., Kasprzak, A.: Task Allocation in Mesh Connected Processors with Local Search Meta-heuristic Algorithms. In: KES-AMSTA 2009. LNCS (LNAI), vol. 5559, pp. 215–224. Springer, Heidelberg (2010)
15. Koszalka, L., Lisowski, D., Pozniak-Koszalka, I.: Comparison of allocation algorithms for mesh structured networks with using multistage simulation. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, pp. 58–67. Springer, Heidelberg (2006)
16. Kasprzak, A.: Packet Switching Wide Area Networks. In: WPWR, Wroclaw (2001) (in Polish)

An Experimentation System for Testing Bee Behavior Based Algorithm to Solving a Transportation Problem

Adam Kakol[1], Iwona Pozniak-Koszalka[1], Leszek Koszalka[1], Andrzej Kasprzak[1], and Keith J. Burnham[2]

[1] Dept. of Systems and Computer Networks, Wroclaw University of Technology, 50-370 Wroclaw, Poland
[2] Control Theory and Applications Centre, Coventry University, CV1 5FB Coventry, United Kingdom
leszek.koszalka@pwr.wroc.pl

Abstract. This paper describes a system, called eTransport to solve a transportation problem. The system was applied in experiments that test meta-heuristic algorithms for solving such optimization task. The engine of this system is based on a bee behavior based algorithm, called KAPI, designed by the authors. In order to make a comparison of the results obtained by KAPI, the Ants Colony System and Tabu Search algorithms were also applied. The eTransport simulator can generate real-life scenarios on Google Maps, while configuring and running tested algorithms, and finally, displaying the solutions found. Some illustrative examples of experiments are presented and discussed. The analysis of results of multi-scenario simulations shows the advantages of the KAPI algorithm, and justifies the conclusion that KAPI is much better and more effective at finding solutions to problems of this kind than other known algorithms.

Keywords: transportation problem, evolutionary algorithm, experimentation system.

1 Introduction

This paper focuses on a transportation problem described in [1] and relevant to logistics companies. Our formulation of this problem differs slightly from the Modified Vehicle Routing Problem (MVRP) defined in [2] and [3], by taking into consideration that: (i), the cost of production can vary depending on the factory, (ii) each factory is able to use vehicles (trucks) of different types (with different capacities), (iii) every single vehicle may deliver goods not only to one customer. The total cost defined to characterize the quality of a solution, is strongly related to the chosen route and the types of vehicles (trucks) to be used.

In the paper, an improved algorithm based on bee behavior (its first draft was presented in [4]) has been evaluated by comparing with other meta-heuristic algorithms, including an implementation of the Ants Colony System (ACS) – a description of ACS ideas may be found in [5], [6], [7], and [8], and an implementation of Tabu Search algorithm (TS) - a description of ACS ideas may be found in [4], [6], and [9].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 11–20, 2011.

Basic concepts and mathematical aspects of the problem are briefly described in Section 2. A concise description of algorithms used to solve a transportation problem is given in Section 3. A description of the experimentation system can be found in Section 4., and a design of the experiments in Section 5. Results of investigations are presented in Section 6. Some conclusions and final remarks appear in Section 7. Suggestions for the further research in the considered area are given in Section 8.

2 Problem Statement

A transportation problem can be described in the following way: a logistics company has [M] factories supplying products to [O] customers in a certain area. Each factory produces, depending on its size, homogeneous goods [T] to be delivered to recipients according to their demands [Z]. In each factory, there is a specified quantity of vehicles [S] of different types [R]. The cost of production [K] of each goods item depends on the factory where it was made. The cost of transportation depends on the distance of the routes [D], the size and load level of vehicle [Ps], and the driver's salary [W]. The following constraints are taken into consideration:

- in each factory, there must be sufficient number of vehicles to be able to take out the whole production from warehouse:

$$T_{total} \leq \sum_{i=1}^{S} P_{si} \tag{1}$$

where: P_{si}-capacity of i-th truck, S-the total number of trucks, T_{total} -the total quantity of goods;

- quantity of products is equal or greater than the actual demand

$$\sum_{i=1}^{F} T_i \quad \geq \sum_{j=1}^{O} Z_j \tag{2}$$

where: Ti - quantity of goods in i- th factory, Zj-demand for the goods by the recipient, O-the number of customers.

The total cost is expressed by formula (3):

$$K_c = \sum_{i=1}^{F} K_{w_i} * I_{w_i} + \sum_{j=1}^{S_{used}} \sum_{k=1}^{I_T} K_{Tj}(D, Z, R) + W \tag{3}$$

where: K_{wi}-the cost of production in i-th factory, I_w-the quantity of exported goods from i-th factory, S_{used}-the number of vehicles to be used for transport, I_T-the number of paths travelled byvehicle between the origin and destination nodes, D – the distance ridden by j-th vehicle, Z_s- the vehicle load level, R-the capacity of vehicle, W - driver's salary (with fixed value), F-the number of factories.

The considered transportation problem consists in minimization (3), i.e. determining the load of any vehicle and its route such that the total cost of transportation goods from factories to customers is minimum.

For the purposes of this paper, the problem's input data is a map in the form of a graph, which consists of distinct points (nodes) in which are located customers and factories. Every single link between two nodes has a specified known length. Factories, in which goods are produced, have a defined cost of production, a number of vehicles and a quantity of goods in warehouses. Vehicles have known but possibly different load capacities.

3 Algorithms

Ant Colony System (ACS). The detailed description of the idea and the classic version of algorithm based on ant behavior are presented in details in [8], [9]. Our implementation of ACS consists of three stages: (i) the path stage with finding the shortest path from factories to customers, (ii) the auction stage, at which ants (representing a type of vehicles) declare the estimated costs of transport to the chosen customer (recipient) and, all tasks are allocated to ants, (iii) the final stage consisting in finding the best way to the customer for ants selected at the auction stage.

The input parameters for ACS are: the number of cycles for stages, the quantity of deposited pheromone; the pheromone evaporation rate; and a factor which represents the strength of the impact of the pheromone.

Tabu Search (TS). The specific of the considered problem caused that adaptation of Tabu Search concept (see e.g. [4], [9]) required making some modifications in classic version of this algorithm. The basic modification consists in creating two collections of data (variable and fixed data). The fixed collection contains all customers while the variable collection contains all vehicles from all the factories. The order of vehicles in the variable collection defines a solution. The vehicle on the first position in the variable collection provides the customers as long as it has goods on board. Then, next vehicle are sent, supplying other customers in the same way, until satisfying demands of all customers. The shortest route between the factory and the customers is to be found by applying well-known Dijkstra algorithm [15]. To find a new member of neighborhood the successive trucks from the variable collection are swapped. The function for evaluating the quality of solutions is the total cost expressed by (3). The parameters that may be adjusted and may control the performance of the implemented TS algorithm are: the size of tabu list, and the number of cycles.

Conceptual Algorithm based on Bees Intelligence (KAPI). The idea of a KAPI algorithm has its origin in [10] which describes the behavior of a bees' swarm when they look for the best place for a settlement. Special bees are selected from swarm, and fly out in different directions looking for the best location to make a wild beehive. After finding it, they go back to the rest of the bees and perform a specific dance. Energy and commitment are features of the dance: the extent to which these features are displayed in the dance is proportional to the level of satisfaction with the discovered location. The more bees that dance vigorously, pointing in one direction, the greater are the chances that the whole swarm would decide to go there. It was observed that just fifteen bees are needed to persuade the whole swarm acceptance of the location of hive.

In order to solve the MVRP, the bee-hive methodology was adapted [11]. It may be distinct several stages, which are shown in Fig. 1.

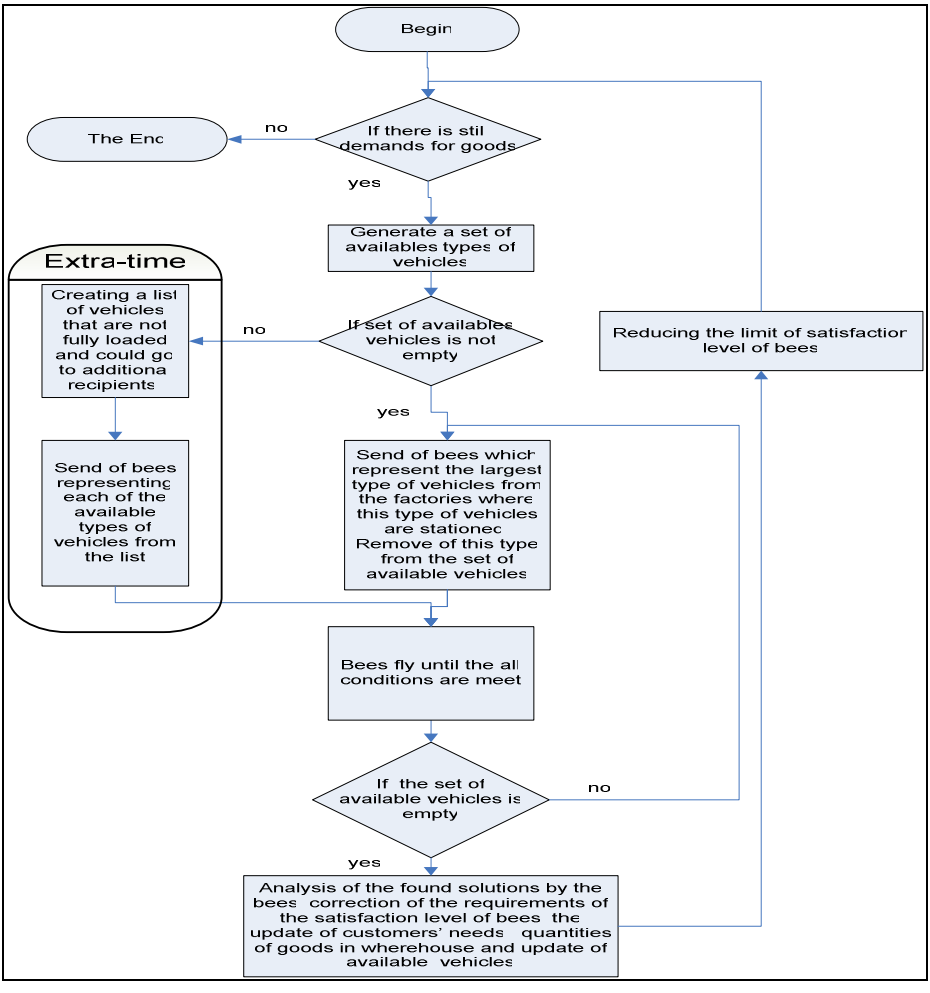


Fig. 1. Block-diagram of modified KAPI algorithm

First, the types of available vehicles are defined. Bees are sent to each of these types. They go carrying a virtual basket that represents the capacity of the vehicle. Each basket is filled with virtual goods, which represent a homogeneous product that can be transported by a vehicle Bees begin their flight from factories, in which there are available the selected types of vehicles. At first, bees are starting with baskets which represent the recently largest-capacity vehicles. Bees fly until the moment when in each recipient node there will be a determined number of bees. At this point, bees with slightly smaller baskets set out on their journey. These steps are repeated until the all available load capacities of the vehicles are exhausted. After the flights,

an analysis of the proposed solutions begins. Each bee has to determine how satisfied it is with the solutions it has found. As a measure of bee satisfaction, we take the cost of sending vehicle to the given recipient using the route founded.

At each customer node, the average satisfaction level is computed. Analysis begins at the place where this average value is the lowest. The analysis consists in answering the following questions (checking conditions):

- whether a bee is sufficiently satisfied (the satisfaction level has to be higher than the minimum required),
- whether there is still demand for supplied goods,
- whether in the factory, from which the bee has begun its flight, there is a sufficient quantity of goods,

If all these conditions are met, an update is made of customers' demands; of the number of goods suppliers in warehouses; and concerning vehicle availability. If no complete solution has been found by this stage, all the previous steps are repeated, but only after firstly reducing the minimum level of required satisfaction. The final solution consists of the chosen vehicles, the specified route for each vehicle, and the determined set of customers which have to be visited.

4 Experimentation System

A new simulation system, called eTransport, was designed and implemented in order to test the performances of meta-heuristic algorithms dedicated to solving a transportation problem. The user interface of eTransport is shown in Fig. 2. The eTransport system is implemented via technologies used to create Web applications, i.e. HTML, CSS, PHP, MySql, Javascript and AJAX. [12].

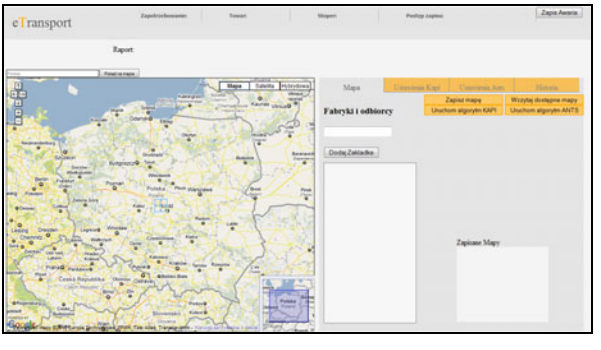


Fig. 2. The interface of eTransport simulator

The application is divided into two autonomous parts. The first is the engine, which consists of two algorithms (KAPI and ACS), written in PHP language. The second part is a component responsible for data presentation. It consists of several parts, exploiting *html* with cascading style sheets, maps supported by Google maps API [13]. The control mechanism is written in JavaScript and AJAX. The following input data defines MVRP: (i) Localization of customers and factories, (ii) Number and type of vehicles, (iii) Quantity and the cost of goods in each factory, (iv) Customer demands.

5 Design of Experiment

The aim of this study was to determine which of the considered algorithms is better at solving the MVRP. Algorithms were tested in many different situations, for different input parameters. The main criteria for evaluating the quality of results was the total cost defined by (3), and the computational time needed to find the solution. The investigations consisted of a series of experiments which were carried out on various types of maps (notion 'map' is defined in [4]) created specially for the purpose of this work. Maps differ in the number of factories and customers, in the distances between them, in the cost of production, and in the ratio of factories to customers. The complex experiment was divided into two parts: Parts I for comparison of KAPI and ACS and Part II for comparison of KAPI and TS.

In Part I, three scenarios (Case 1, Case 2, and Case 3) were designed and examined. The scenarios differ in the quantity and type of vehicles available, and in the ratio between the number of demanded goods and the number of available goods. Such experiment design has followed idea of multistage experiments [14]. Both algorithms, KAPI and ACS were activated for the same designs of experiments.

In Part II, input problem parameters were the same for KAPI and TS, including the number of factories equal to 10 and the number of customers equal to 10, either. The different algorithm's parameters were taken into consideration. Each series of experiment consisted of 10 single simulations.

6 Investigations

Case 1. The map contains 12 distinct locations. The number of factories is the same as the number of customers. The company has its factories in a relatively small area. It is assumed that the factories produce goods with the averaged cost of 2380 Polish currency.

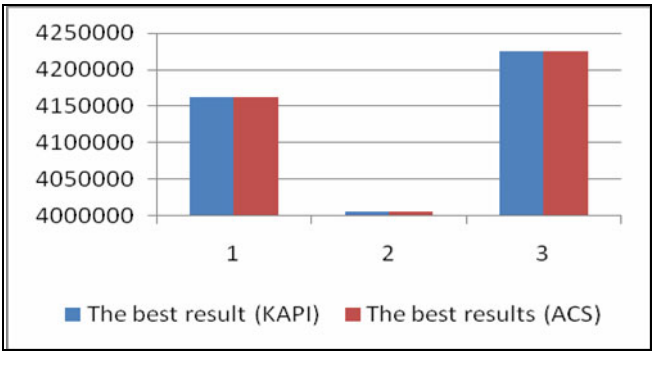


Fig. 3. Results of KAPI and ACS for three scenarios of map 12

Tests performed in Case 1 show that ACS achieves similar results (Fig. 3) with shorter time of execution (Fig. 4).

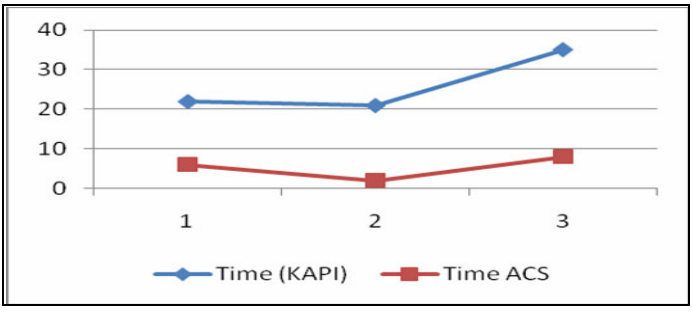


Fig. 4. Time of generating the best results [s] for three scenarios (map 12)

Case 2. The company in Spain has a few factories, but the market is much larger. Distances between points on the map are very long. The average distance between points is more than 500km. The unit price of goods is low: it can be assumed as 2.5 euro. There are no great concentrations of users. In Spain, roads are organized as a cobweb. All the main roads run toward a central point, which is Madrid.

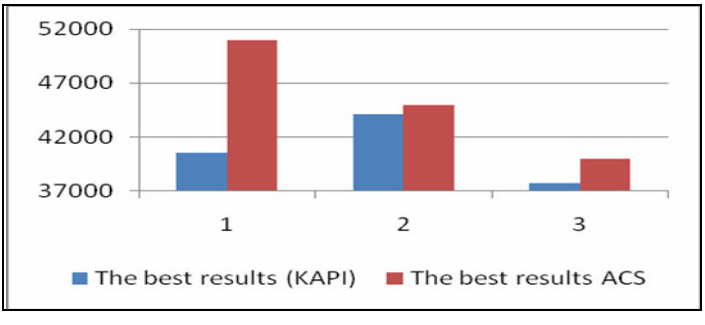


Fig. 5. Comparison KAPI and ACS (map 29)

Now, it can be observed that KAPI works better. It gives better results in comparison to ACS (Fig. 5). The computational time depends on the scenario.

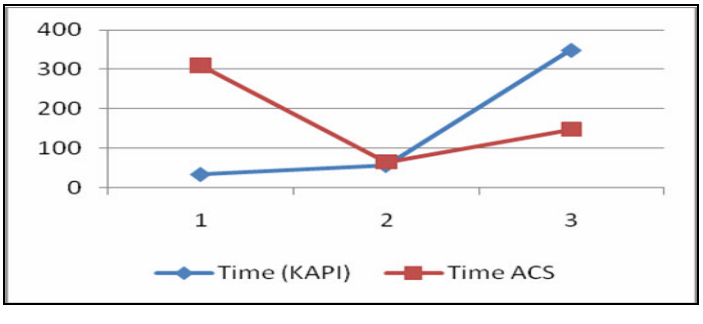


Fig. 6. Time of generating the best results [s] for three scenarios (map 29)

Solving the first scenario with KAPI took 4 times less time than using ACS (Fig. 6).

Case 3. The company has a very large market. On the map there were 60 placed points, (55 customers and 5 factories). Customers' nodes are located in the 55 major cities in Poland (almost 25 percent of them is located in the Upper Silesia region). The average distance between points on the map is less than it was in previous case (in this case it is less than 320 km). The average cost of production was set at 2.7. Factories were placed in large cities, which were selected to minimize their distances to the largest possible groups of customers. The road network in the third map is rather regular. It does not exhibit any characteristic structure.

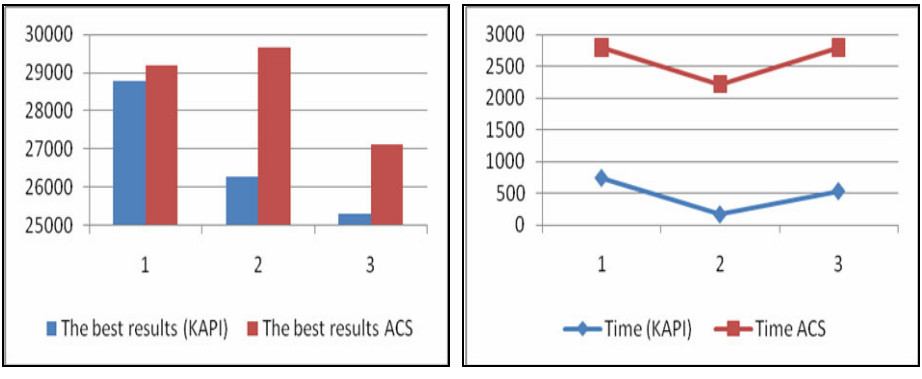


Fig. 7. Results of KAPI and ACS algorithm for the map 60, (a) cost, (b) time

Analyzing the obtained results, it may be observed, that KAPI worked very well. Both, the costs expressed by the formula (4) and shown in Fig. 7 (a), and the computational times (Fig. 7 (b)) are much better in comparison to ACS algorithm.

Summary. Fig. 8. presents the best solutions found by the KAPI and ACS for each of the investigated maps (denoted by 12, 29, 60). The results are expressed as a percentage instead of an average. For map 12 the results are more or less the same. It is due to the fact that the cost of a unit of goods is very high, and it has the greatest impact on the final cost (the transport cost is very low, relatively).

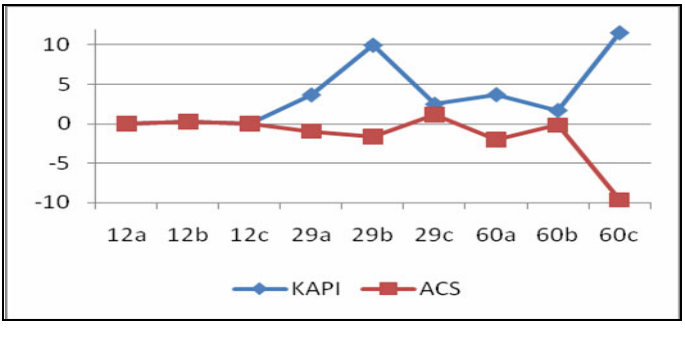


Fig. 8. Comparison of KAPI and ACS [%] for various cases and scenarios

Discussion. The study shows that the KAPI algorithm is more effective in almost all of the tested scenarios. ACS performed slightly better when solving map 12, in particular in a special situation where the number of factories is equal to the number of customers (recipients). In the situations where there are many vehicles of the same type, KAPI can solve the problem much faster than the ACS. Once the company has plenty of vehicles of different load capacities, then it is better to use ACS.

Case 4. The obtained results are shown in special format where points represent the average 'score' for series of experiments while lower ends of lines, which are coming from those points, show the best found result.

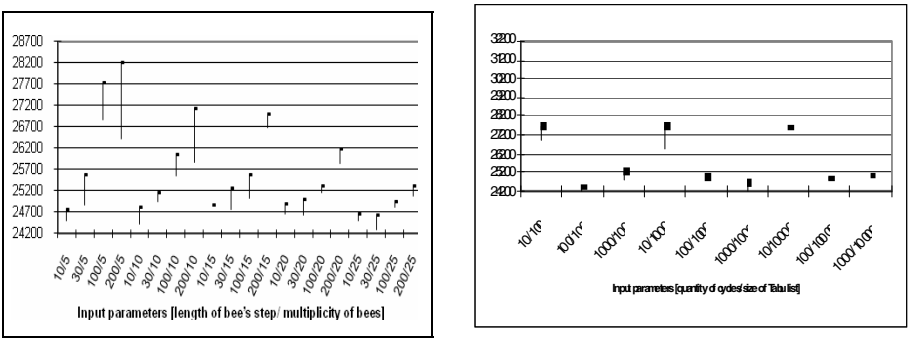


Fig. 9. Results of (a) KAPI algorithm and (b) TS algorithm

Summary. We can already spot the big difference in results given by KAPI algorithm just by changing its parameters. Analyzing Fig. 9 (a) it can be easily concluded that increasing the length of step adversely (KAPI parameter) affects the quality of solutions. For this map best solutions were obtained for the 'the maximum number of bees in the node' (KAPI parameter) set to 5 or 15. The best solution was found with the parameter set to 15.

Discussion. In theory, TS should return the best result when running with high value of parameters (e.g. a large number of cycles as well as the size of a tabu list), yet in this case the best result was found with the number of cycles set to 100 and the same size of tabu list (Fig. 9 (b)).

7 Conclusions

In this paper, we have focused on a some category of transportation problem. For this purpose, a new version of KAPI algorithm was proposed by the authors. The KAPI seems to be very promising, e.g. when comparing KAPI with ACS, we obtained better solution for KAPI in 7 series out of 9 series of experiments.

The designed and implemented experimentation system, in particular its module eTransport simulator, may be used for aiding the solution of the considered category of transportation problem. Moreover, it is possible to visualize every individual solution, thus the user can estimate, really quickly, how good the solution is. A particularly important aspect of this experimentation system is the possibility to work on

real-life problems. Recently, the system is serving as a tool to aid teaching students in control engineering, computer science and telecommunications areas in Wroclaw University of Technology.

8 Perspectives

On the basis of the results of experiments, the authors suggest some improvements such as the following could be made:

- Limitation of the length of a route which a driver can choose,
- Introduction another cost functions [15] (the current version of eTransport relies on a predetermined formula for calculating the cost of transportation).
- More features of the vehicles can be taken into consideration (e.g. fuel combustion, maximum speed, etc).
- Flexible definition of drivers' salary.

References

1. Gendreau, M., Iori, M., Laporte, G.: A Tabu Search Heuristic for the Vehicle Routing Problem with Two-dimensional Loading Constraints. Networks 51, 4–18 (2008)
2. Wagner, T.: Principles of Operations Research with Applications to Managerial Decisions, 2nd edn., Phi Learning (2009)
3. Mathirajan, M., Meenakshi, B.: Experimental Analysis of Some Variants of Vogel's Approximation Method. Indian Institute of Science (2003)
4. Kakol, A., Grotowski, K., Koszalka, L., Pozniak-Koszalka, I., Kasprzak, A., Burnham, K.J.: Performance of Bee Behavior-Based Algorithm for Solving Transportation Problem. In: Proc. of 5th ICONS IARIA Conference, pp. 83–87 (2010)
5. Bullnheimer, Hartl, R.F., Strauss, C.: An Improved Ant System Algorithm for the Vehicle Routing Problem. Technical Report POM-10/97, Institute of Management Science, University of Vienna, Austria (1997)
6. Dorigo, M., Di Caro, G., Gambardella, L.M.: Algorithms for Discrete Optimization. University Libre de Bruxelles, Belgium (2003)
7. Uso del algoritmo de optimización de colonia de hormigas. Investigación e Innovación para Ingeniería Civil (2007) (in Spanish), http://www.peru-v.com/
8. Gutjahr, W.J.: A Graph–Based Ant System and Its Convergence. Future Generation Computing System 16, 873–888 (2000)
9. Abounacer, R., Bencheikh, G., Boukachour, J., Dkhissi, B., Alaoui, A.E.: Population Meta- heuristics to Solve the Professional Staff Transportation Problem. CERENE Laboratory, ISEL, Quai Frissard B.P. 1137 76063 Le Havre, Cedex, France (2003)
10. Miller, P.: Theory of Swarm. National Geographic (August 2007)
11. Kakol, A., Grotowski, K., Łobodziński, T., Koszalka, L., Pozniak-Koszalka, I.: CABI Algorithm for Solving Unbalanced Transportation Problem. In: Proc. of ICSE, Coventry, UK (2009)
12. Hayder, H.: Object Programming with PHP 5. Helion, Gliwice (2009) (in Polish)
13. Koszalka, L., Lisowski, D., Pozniak-Koszalka, I.: Comparison of Allocation Algorithms for Mesh Structured Networks Using Multistage Simulation. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, pp. 58–67. Springer, Heidelberg (2006)
14. http://code.google.com/intl/pl/apis/maps/
15. Kasprzak, A.: Packet Switching Wide Area Networks. In: WPWR, Wroclaw (2001) (in Polish)

Multi-response Variable Optimization in Sensor Drift Monitoring System Using Support Vector Regression

In-Yong Seo, Bok-Nam Ha, and Min-Ho Park

KEPCO Research Institute,
65, Munji-Ro, Yuseong-Gu, Daejeon, Korea
iyseo@kepri.re.kr

Abstract. In a nuclear power plant (NPP), periodic sensor calibrations are required to assure sensors are operating correctly. However, only a few faulty sensors are found to be calibrated. For the safe operation of an NPP and the reduction of unnecessary calibration, on-line calibration monitoring is needed. Most researches have been focused on improving only the accuracy of the system although sensitivity is another important performance index. This paper presents multi-response optimization for an on-line sensor drift monitoring system to detect drift and estimate sensor signal effectively. Accuracy and sensitivity of the principal component-based auto-associative support vector regression (PCSVR) were optimized at the same time by desirability function approach. Response surface methodology (RSM) is employed to efficiently determine the optimal values of SVR hyperparameters. The proposed optimization method was confirmed with actual plant data of Kori NPP Unit 3. The results show the trade-off between the accuracy and sensitivity of the model as we expected.

Keywords: PCA, SVR, Multi-Response, RSM, NPP.

1 Introduction

For the past two decades, the nuclear industry has attempted to move toward a condition-based maintenance philosophy using new technologies developed to monitor the condition of plant equipment during operation. The traditional periodic maintenance method can lead to equipment damage, incorrect calibrations due to adjustments made under non-service conditions, increased radiation exposure to maintenance personnel, and possibly, increased downtime. In fact, recent studies have shown that less than 5% of the process instruments are found in degraded condition that require maintenance. Therefore, plant operators are interested in finding ways to monitor sensor performance during operation and to manually calibrate only the sensors that require correction.

Considerable research efforts have been made to develop on-line calibration monitoring algorithms. The application of artificial intelligence techniques to nuclear power plants were investigated for instrument condition monitoring [1]. The Multivariate State Estimation Technique (MSET) was developed in the late 1980s [2], and the Plant Evaluation and Analysis by Neural Operators (PEANO) was developed in [3] by using auto-associative neural networks. The SVR algorithm that was developed

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 21–30, 2011.

by Vapnik [4] is based on the statistical learning theory. The SVM method was applied for the data-based state estimation in nuclear power reactors [5].

The use of support vector regression (SVR) for on-line monitoring and signal validation was developed by Korea Electric Power Research Institute (KEPRI) and reported in NUCLEAR ENGINEERING AND TECHNOLOGY [6, 7]. The research presented in this paper primarily focuses on multi-response optimization for an on-line sensor drift monitoring. Accuracy and sensitivity of the PCSVR model were optimized at the same time by desirability function approach. The proposed optimization method was confirmed with actual plant data of Kori NPP Unit 3.

2 PC- Based AASVR

The outputs of an auto-associative model are trained to emulate its inputs over an appropriate dynamic range. An auto-associative model will estimate the correct input values using the correlations embedded in the model during its training. The estimated correct value from the auto-associative model can then be compared to the actual process parameter to determine if a sensor has drifted or has been degraded by another fault type. Fig. 1 shows the schematic diagram of the proposed PCSVR method for modeling measurements in an NPP.

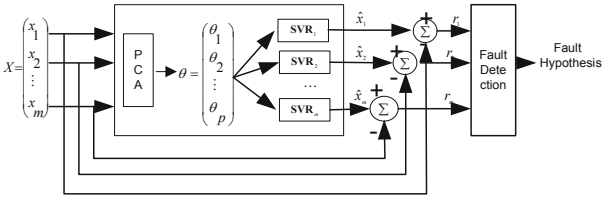


Fig. 1. The schematic diagram of PCSVR

2.1 AASVR Based upon Principal Components

In this paper, an SVM regression method is used for signal validation of the measurements in NPPs. The SVM regression is to nonlinearly map the original data into a higher dimensional feature space. Hence, given a set of data $\{(X_i, y_i)\}_{i=1}^{n} \in R^m \times R^m$ where $\mathbf{x}_i$ is the input vector to support vector machines, $\mathbf{y}_i$ is the actual output vector and n is the total number of data patterns. The multivariate regression function for each output signal is approximated by the following function,

$$y_k = f_k(\boldsymbol{x}) = \boldsymbol{w}_k{}^T \phi(\boldsymbol{x}) + b_k \ . \tag{1}$$

where $\boldsymbol{w}_k = [w_1, w_2, \cdots, w_n]^T$, $\phi = [\phi_1, \phi_2, \cdots, \phi_n]^T$, $k = 1,2,\cdots,m$ and m is the number of sensor measurements. Also, the function $\phi_i(\boldsymbol{x})$ is called the feature. Equation (1) is a

nonlinear regression model because the resulting hyper-surface is a nonlinear surface hanging over the *m*-dimensional input space. The parameters **w** and *b* are a support vector weight and a bias that are calculated by minimizing the following regularized risk function:

$$R(\boldsymbol{w}_k) = \frac{1}{2}\boldsymbol{w}_k{}^T\boldsymbol{w}_k + C_k\sum_{i=1}^{n} L_k(y_{k,i}) \cdot \tag{2}$$

where

$$L_k(y_{k,i}) = \begin{cases} 0, & \left|y_{k,i} - f_k(\boldsymbol{x})\right| < \varepsilon_k \\ \left|y_{k,i} - f_k(\boldsymbol{x})\right| - \varepsilon_k, & otherwise \end{cases} \tag{3}$$

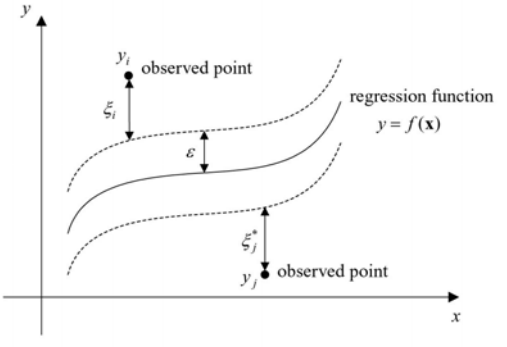


Fig. 2. The Parameters for the AASVR Models

The first term of Equation (2) characterizes the complexity of the SVR models. C_k and ε_k are user-specified parameters and $L_k(y_{k,i})$ is called the ε-insensitive loss function [8]. The loss equals zero if the estimated value is within an error level, and for all other estimated points outside the error level, the loss is equal to the magnitude of the difference between the estimated value and the error level. That is, minimizing the regularized risk function is equivalent to minimizing the following constrained risk function:

$$\text{Minimize} \quad R(\boldsymbol{w}, \xi, \xi^*) = 1/2\boldsymbol{w}_k{}^T\boldsymbol{w}_k + C_k\sum_{i=1}^{n}(\xi_{k,i} + \xi_{k,i}^*) \tag{4}$$

$$\text{Subject to} \quad \begin{aligned} & y_{k,i} - \mathbf{w}_k^T\phi(\mathbf{x}) - b_k \le \varepsilon_k + \xi_{k,i} \\ & \mathbf{w}_k^T\phi(\mathbf{x}) + b_k - y_{k,i} \le \varepsilon_k + \xi_{k,i}^* \\ & \varepsilon_k, \xi_{k,i}, \xi_{k,i}^* \ge 0 \quad for\ i = 1,2,\cdots,n \end{aligned} \tag{5}$$

where the constant C determines the trade-off between the flatness of $f(\boldsymbol{x})$ and the amount up to which deviations larger than ε are tolerated, and ξ and ξ^* are slack variables representing upper and lower constraints on the outputs of the system and are positive values.

The constrained optimization problem can be solved by applying the Lagrange multiplier technique to (4) and (5), and then by using a standard quadratic programming technique. Finally, the regression function of (1) becomes

$$y_k = \sum_{i=1}^{n} (\lambda_{k,i} - \lambda_{k,i}^*) \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b_k^* \tag{6}$$

where $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x})$ is called the kernel function.

By using different kernel functions for inner product evaluations, various types of nonlinear models in the original space could be constructed. It has been shown that, in general, radial-basis function (RBF) is a reasonable first choice of kernel functions since it equips with more flexibility and less parameters. The RBF kernel function used in this paper is expressed as

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{ -\frac{(\mathbf{x} - \mathbf{x}_i)^T (\mathbf{x} - \mathbf{x}_i)}{2\sigma^2} \right\} \tag{7}$$

where σ is the kernel function parameter. The bias, b, is calculated as follows:

$$b_k^* = -\frac{1}{2} \sum_{i=1}^{n} (\lambda_{k,i} - \lambda_{k,i}^*)[K(\mathbf{x}_r, x_i) + K(\mathbf{x}_s, x_i)] \cdot \tag{8}$$

where $\mathbf{x}_r$ and $\mathbf{x}_s$ are called support vectors (SVs) and are data points positioned at the boundary of the ε-insensitivity zone. By replacing principal component $\boldsymbol{\theta}$ with **x,** we can combine PC and AASVR as follows:

$$\begin{aligned} y_k &= f_k(\theta) = \sum_{i=1}^{n} (\lambda_{k,i} - \lambda_{k,i}^*) \boldsymbol{K}(\boldsymbol{\theta}_i, \boldsymbol{\theta}) + b_k^* \cdot \\ b_k^* &= -\frac{1}{2} \sum_{i=1}^{n} (\lambda_{k,i} - \lambda_{k,i}^*)[K(\theta_r, \theta_i) + K(\theta_s, \theta_i)] \end{aligned} \tag{9}$$

The three most relevant design parameters for the AASVR model are the insensitivity zone, ε, the regularization parameter, C, and the kernel function parameter, σ. An increase in the insensitivity zone, ε, reduces the accuracy requirements of the approximation and allows a decrease in the number of SVs. In addition, an increase in the regularization parameter, C, reduces larger errors, thereby minimizing the approximation error. The kernel function parameter, σ, determines the sharpness of the radial basis kernel function.

2.2 Desirability Function Approach for the Multi-response Optimization

Experimental design, especially Response Surface Methodology (RSM) is widely used in off-line quality improvement. For most products and processes, quality is

multi-dimensional, it's common to observe multiple responses on experimental units and optimize these responses simultaneously. And the desirability function approach is widely used in practice and in many commercial software packages such as Minitab and JMP.

This experiment was performed with PCSVR model using the NPP operation data. A central composite design is used with two independent (or uncorrelated) responses: accuracy (y_1), and auto-sensitivity (y_2). For the auto-sensitivity experiments we added artificial drift to each sensor signal which linearly increases with time from 0 % to 4 % of its nominal value at the end point. The goal here is to determine the optimal setting of sigma (x_1), epsilon (x_2), and C (x_3) to minimize the y_1 and y_2 simultaneously. Based on the engineering requirements, y_1 which is measured by the natural logarithm of MSE should be in the range of -5.0 to -3.5 and y_2 in 0.3 to 0.45. The data of this experiment are shown in Table 1.

Table 1. Central composite design with three variables and two responses

Standard Order	Uncoded Variables			Responses	
	Sigma (x_1)	Epsilon (x_2)	C (x_3)	Accuracy (y_1)	Auto Sensitivity (y_2)
1	0.5649	0.0105	2.1067	-4.1174	0.4598
2	1.6351	0.0105	2.1067	-4.3767	0.5266
3	0.5649	0.0400	2.1067	-3.2843	0.3019
4	1.6351	0.0400	2.1067	-3.3081	0.4038
5	0.5649	0.0105	7.9933	-4.1393	0.4660
6	1.6351	0.0105	7.9933	-4.4404	0.5395
7	0.5649	0.0400	7.9933	-3.2843	0.3019
8	1.6351	0.0400	7.9933	-3.3108	0.4142
9	0.2000	0.0253	5.0500	-2.9201	0.3437
10	2.0000	0.0253	5.0500	-3.6982	0.4832
11	1.1000	0.0005	5.0500	-5.0443	0.5902
12	1.1000	0.0500	5.0500	-3.1095	0.3408
13	1.1000	0.0253	0.1000	-3.3083	0.3077
14	1.1000	0.0253	10.000	-3.6820	0.4221
15	1.1000	0.0253	5.0500	-3.6820	0.4221
16	1.1000	0.0253	5.0500	-3.6888	0.4224
17	1.1000	0.0253	5.0500	-3.6669	0.4218

Simultaneous consideration of several responses involves first building an appropriate response surface model for each response. The fitted full second-order response surface models for the two responses are given respectively as follows:

$$y_1 = -4.213 - 0.997x_1 + 76.5315x_2 - 0.0526x_3 + 0.3338x_1^2 - 811.96x_2^2 + 0.00345x_3^2$$

$$y_2 = 0.430 + 0.08275x_1 - 9.1534x_2 + 0.0258x_3 - 0.00097x_1^2 + 83.762x_2^2 - 0.00201x_3^2$$

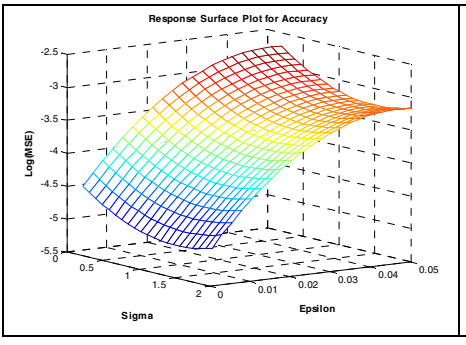


Fig. 3. Response surface plot of accuracy

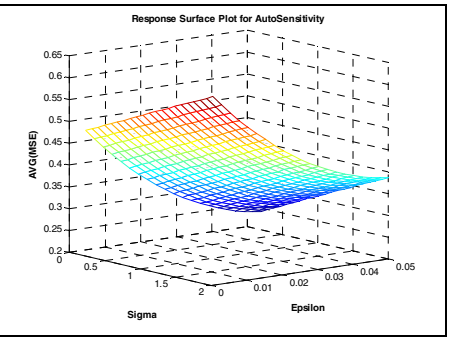


Fig. 4. Response surface plot of auto sensitivity

To calculate individual desirability for each response, we employ traditional desirability function method proposed by Derringer and Suich [9]. Let d_i be the ith individual desirability function, L_i and U_i be the lower and upper specification limit respectively. If the target T_i for the response y_i a minimum value (smaller the- better),

$$d_i = \begin{cases} 1 & y_i < T_i \\ \left(\dfrac{U_i - y_i}{U_i - T_i} \right)^r & T_i \le y_i \le U_i \\ 0 & y_i > T_i \end{cases} \tag{10}$$

where r denotes the weight, the change of the value of r will influence the individual desirability function.

Here, we assumed there are no interactions between control variables. To make use of desirability function, we might formulate this problem as

$$D = \left(\prod_{i=1}^{m} d_i^{w_i} \right)^{1/W} \tag{11}$$

where D represents composite desirability or overall desirability, d_i is the individual desirability for the i^{th} response, w_i is the importance of the i^{th} response, and $W = \sum_{i=1}^{m} w_i$ m is the number of responses.

Because the multi-response optimization highly depends on the importance ratio, we tested the desirability at 10 importance ratio (w_1/w_2) points as shown in Fig.5. The desirability for auto-sensitivity (w_2) was fixed to 1 and the desirability for accuracy (w_1) was changed from 0.1 to 10. From the Fig. 5 we determined the importance ratio of 4 for the optimal solution by engineering knowledge, which can improve the sensitivity while not decreasing the accuracy much.

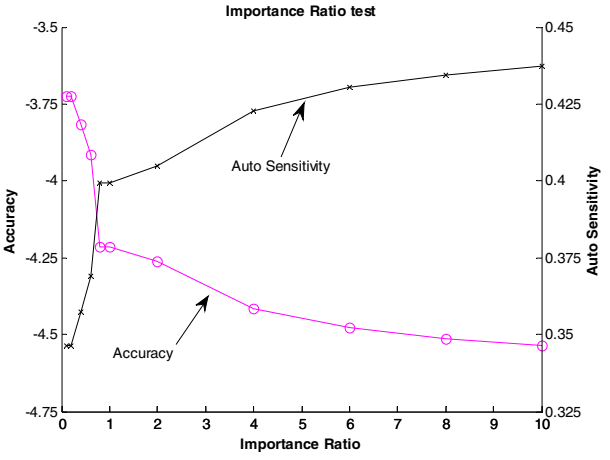


Fig. 5. Accuracy and auto-sensitivity in terms of Importance Ratio

Fig. 6 shows the response surface of desirability function when importance ratio is 4. We can notice that the optimal solution is

$$(\sigma, \varepsilon, C) = (2.0,\ 0.0005,\ 2.0) \tag{12}$$

at which the composite desirability becomes maximum.

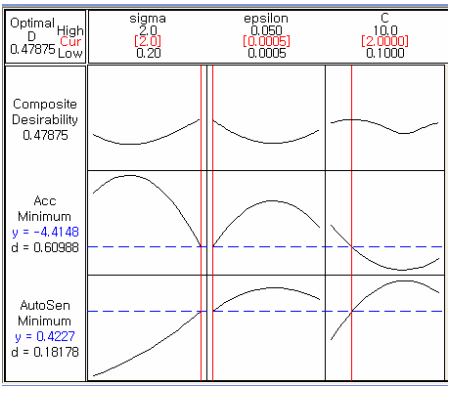


Fig. 6. Response surface of Desirability Function

3 Application to the NPP Measurements

3.1. Experimental Data

The proposed algorithm was confirmed with the real plant startup data of the Kori Nuclear Power Plant Unit 3. These data are the values measured from the primary and

secondary systems of the NPP. The data is derived from the following 11 types of measured signals: the reactor power (the ex-core neutron detector signal, Sensor 1); the pressurizer water level (Sensor 2); the SG steam flow rate (Sensor 3); the SG narrow range level (Sensor 4); the SG pressure (Sensor 5); the SG wide-range level (Sensor 6); the SG main feedwater flow rate (Sensor 7); the turbine power (Sensor 8); the charging flow rate (Sensor 9); residual heat removal flow rate (Sensor 10); and the reactor head coolant temperature (Sensor 11).

The data were sampled at a rate of 1 minute for about 38 hours. The total observation number of measurement data is 2,290 and this data set was normalized in each dimension. The data set was divided into five subsets of equal size, i.e., one training subset, one test subset and three optimization subsets. Total data set was indexed using Arabic numerals, i.e., i = 1,2,. . ., 2,290. 458 patterns with the indices, i=5j+3, j=0,1,…,457, named z3 were used to train SVR to capture the quantitative relation between 11 inputs and outputs. z1 which has indices of 5j+1, j=0,1,…,457, used for the test of the model, while the remaining three subsets (z2, z4, z5) for the optimization.

Let $(\theta_1, \theta_2, \cdots, \theta_{11})$ denote principal components (PCs) obtained by applying PCA to the above plant data. As mentioned earlier, variance is used in selecting dominant PCs. We found that θ_1 is the most dominant PC and explains about 84.12% of total variation in the original data. However, in order to minimize loss of information, the first seven PCs are considered in this study. The selected PCs explain more than 99.98% of total variation. The loss of information is less than 0.1%.

3.2 Test Results

Empirical model building is done using PCSVR hyperparameters derived in (12). The numbers of support vectors needed for each SVR are 377 (82.3%), 286 (62.4%), 388 (84.7%), 182 (39.7%), 358 (78.1%), 266 (58.1%), 421 (91.9%), 378 (82.5%), 229 (50.0%), 85 (18.5%), and 421 (91.9%). The average number of support vectors is 308.2 (67.3%) which is a little bigger than that of in [7].

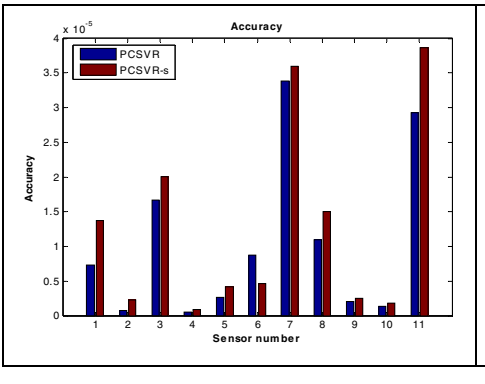


Fig. 7. Accuracy comparison

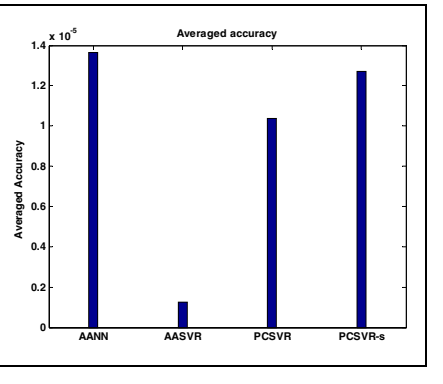


Fig. 8. Averaged accuracy

Fig. 7 shows accuracies of each sensor for the test data. Bars in right-hand side (PCSVR-s) in Fig.7 are accuracies using the multi-response optimization. The averaged accuracies for several models are shown in Fig. 8. From this figure we can notice that the accuracy by using multi-response optimization is a little bit degraded.

In order to test the sensitivity, we artificially degraded the SG main feed water flow rate signal in normalized test data z1 which is shown with a straight line in Fig. 11. Fig. 9 shows sensitivities of each sensor for the test data. Bars in right-hand side in Fig.9 are sensitivities using the multi-response optimization. The averaged sensitivities for several models are shown in Fig. 10. Note that the sensitivity by using multi-response optimization is slightly improved.

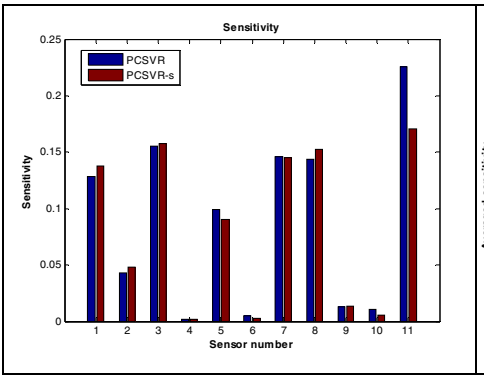


Fig. 9. Sensitivity comparison

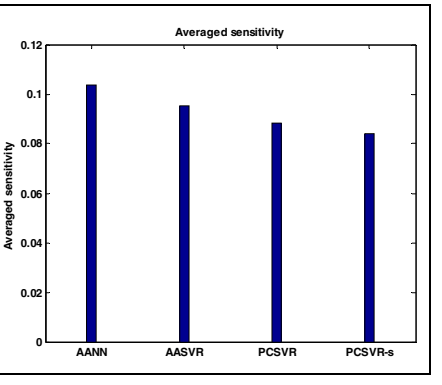


Fig. 10. Averaged sensitivity

Fig. 11 represents residuals of predicted feedwater flow signal when artificially degraded signal was inputted to the PCSVR model which tuned by the multi-response optimization method. From the figure we can know that this model easily detect the sensor drift.

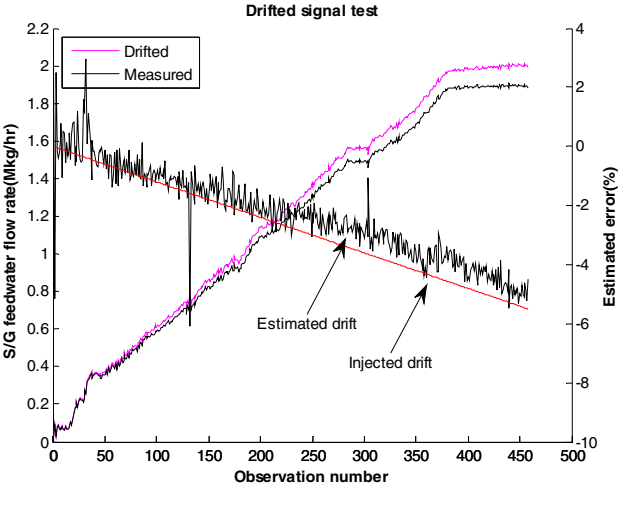


Fig. 11. Detected residuals for the drifted signal

4 Conclusions

In this paper, multi-response optimization for an on-line sensor drift monitoring system is presented to detect drift and estimate of sensor signal effectively. Accuracy and sensitivity of the principal component-based Auto-Associative support vector regression were optimized at the same time by desirability function approach. Response surface methodology is employed to efficiently determine the optimal values of SVR hyperparameters. The proposed optimization method was confirmed with actual plant data of Kori NPP Unit 3. The results show the trade-off between the accuracy and sensitivity of the model as we expected.

References

1. Upadhyaya, B.R., Eryurek, E.: Application of Neural Networks for Sensor Validation and Plant Monitoring. Nuclear Technology 97, 170–176 (1992)
2. Mott, Y., King, R.W.: Pattern Recognition Software for Plant Surveillance, U.S. DOE Report (1987)
3. Fantoni, P., Figedy, S., Racz, A.: A Neuro-Fuzzy Model Applied to Full Range Signal Validation of PWR Nuclear Power Plant Data. In: FLINS 1998, Antwerpen, Belgium (1998)
4. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning 20, 273–297 (1995)
5. Zavaljevski, N., Gross, K.C.: Support Vector Machines for Nuclear Reactor State Estimation. In: ANS International Topical Meeting, Pittsburgh, USA, May 7-11 (2000)
6. Seo, I.-Y., Kim, S.J.: An On-line Monitoring Technique Using Support Vector Regression and Principal Component Analysis. In: CIMCA 2008, Vienna, Austria, December 10-12 (2008)
7. Seo, I.-Y., Ha, B.-N., Lee, S.-W., Shin, C.-H., Kim, S.-J.: Principal Components Based Support Vector Regression Model for On-line Instrument Calibration Monitoring in NPPs. Nucl. Eng. Tech. 42(2), 219–230 (2010)
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
9. Derringer, G., Suich, R.: Simultaneous optimization of several response variables. Journal of Quality Technology 12, 214–219 (1980)

A Method for Scheduling Heterogeneous Multi-Installment Systems

Amin Shokripour, Mohamed Othman*,
Hamidah Ibrahim, and Shamala Subramaniam

Department of Communication Technology and Network,
Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia
shokripour@gmail.com, mothman@fsktm.upm.edu.my

Abstract. During the last decade, the use of parallel and distributed systems has become more common. Dividing data is one of the challenges in this type of systems. Divisible Load Theory (DLT) is one of the proposed method for scheduling data distribution in parallel or distributed systems. Many researches have been done in this field but scheduling a multi-installment heterogeneous system in which communication mode is blocking has not been addressed. In this paper, we present some closed-form formulas for the different steps of scheduling jobs in this type of systems. The results of our experiments show the proposed method gave better performances than the Hsu et al.'s method.

1 Introduction

Different models were investigated in DLT researches [1,2], each of which is made by some assumptions. One installment system with blocking and non-blocking mode communication [3], multi-installment system by non-blocking communication modes [4], system with different processor available time (SDPAT) [5], non-dedicated systems [6], and others are some examples of the investigated models. However no closed-form formula has yet been presented for multi-installment system with blocking mode communication yet.

In this paper, we will try to schedule jobs in a multi-installment system with high communication speed. The proposed method considers all four steps in scheduling a multi-installment system.

The remainder of this paper is organized as detailed below. Section 2 presents related works. Our model and notations are introduced in the third section. In section 4, a proposed method which includes closed-form formula for finding the proper number of processors, for finding the proper number of installments, scheduling internal installments, and scheduling the last installment, is presented. The results of experiments and their analysis are appeared in section 5. In the last section, the conclusion is obtainable.

* The author is also an associate researcher at the Lab of Computational Science and Informatics, Institute of Mathematical Research (INSPEM), Universiti Putra Malaysia.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 31–41, 2011.

2 Related Works

One of the first studies that presented a closed-form formula for scheduling multi-installment system was done by Yang et al. [7]. In this research, size of installments were not equal. They also used non-blocking communication mode. The authors presented a closed-form formula for scheduling jobs in a homogeneous and heterogeneous multi-installment system. After this research, some articles were published for improving this method by using different technique such as parallel communication [8], resizing chunk size of each installment inrespects of to the negative effect of performance prediction errors at the end of execution[7], and attending to other system parameters [9].

Hsu et al. presented a new idea for multi-installment scheduling jobs in a heterogeneous environment in [10]. Their article includes two algorithms, ESCR and SCR. ESCR is an extension of SCR which has some problems for scheduling the last installment. They did not attend to communication and computation overheads and the size of each installment is also independent of job size. Therefore, for large job sizes, the increased number of installments results in making some problems for systems with overhead.

3 Preliminaries

Throughout this paper, the following notations and their definitions are used and stated in Table 1.

In this research, we used client-server topology for network, while all processors are connected to a root called P_0. The root does not do any computation and only schedules tasks and distributes chunks among workers. Communication type is blocking mode; it means communication and computation cannot be overlapped. This model consists of a heterogeneous environment which includes communication and computation overheads.

Table 1. Notations

Notation	Description
W	Total size of data
V	Size of each installment
n	Number of installments
m	Number of processors
α_i	The size of allocated fraction to processor P_i in each internal installment
β_i	The size of allocated fraction to processor P_i in the last installment
w_i	Ratio of the time taken by processor P_i, to compute a given load, to the time taken by a standard processor, to compute the same load
z_i	Ratio of the time taken by link l_i, to communicate a given load, to the time taken by a standard link, to communicate the same load
s_i	Computation overhead for processor P_i
o_i	Communication overhead for processor P_i
$T(\alpha_i)$	The required time for transferring and computing data in processor P_i.

4 Proposed Method

In this research, we assume that the size of all installments is equal and call it V. We claim that the calculated proper number of installment is the optimum. Hence, we cannot change the size of each installment because if we change the size of installments and increase the size, the idle time for each processor between each installment is increased because the summation of transferring data to the previous processors is increased but the computation time this processor is static and finished before the finishing previous processors transferring data. If we decrease the size installments, summation of time transferring data to the previous processor is decreased and when the root should send the data to the current processor, it is busy because the computation time of previous installment is larger than summation of communication time of the other processors.

4.1 Internal Installments Scheduling

In each internal installment, we want to have equal values for the total time taken for communication and computation time for all processors (Fig. 1). In other words, we determine the size of task for each processor so that the total time taken for communication and computation for each processor is equal to the total of communication time of all the other processors. For example, when P_1 finishes its computation, the root is ready to send the data of the next installment to it. As a result,

$$\begin{aligned} \alpha_1 V(z_1 + w_1) + o_1 + s_1 &= \alpha_2 V(z_2 + w_2) + o_2 + s_2 = \ldots \\ &= \alpha_m V(z_m + w_m) + o_m + s_m \end{aligned} \tag{1}$$

$$\alpha_2 = \frac{\alpha_1 V(z_1 + w_1) + o_1 + s_1 - o_2 - s_2}{V(z_2 + w_2)} \tag{2}$$

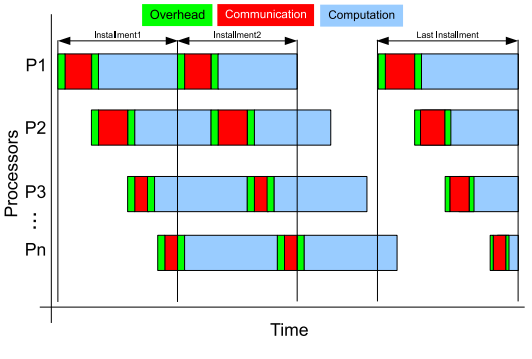


Fig. 1. Time Diagram for a Multi-Installment System

We know that $\alpha_1 + \alpha_2 + \ldots + \alpha_m = 1$. Hence, we have

$$\begin{aligned} \alpha_1 + & \frac{\alpha_1 V(z_1 + w_1) + o_1 + s_1 - o_2 - s_2}{V(z_2 + w_2)} \\ & + \frac{\alpha_1 V(z_1 + w_1) + o_1 + s_1 - o_3 - s_3}{V(z_3 + w_3)} + \ldots \\ & + \frac{\alpha_1 V(z_1 + w_1) + o_1 + s_1 - o_m - s_m}{V(z_m + w_m)} = 1 \end{aligned} \tag{3}$$

Two new variables are defined, $\Delta_i = \frac{z_1 + w_1}{z_i + w_i}$ and $\Phi_i = \frac{o_1 + s_1 - o_i - s_i}{z_i + w_i}$. Now Eq.(3) can be rewritten as

$$\alpha_1 (1 + \sum_{i=2}^{m} \Delta_i) + \frac{1}{V} \sum_{i=2}^{m} \Phi_i = 1 \tag{4}$$

Finally, we find these closed-form formula as

$$\begin{cases} \alpha_1 = \frac{1 - \frac{1}{V} \sum_{i=2}^{m} \Phi_i}{1 + \sum_{i=2}^{m} \Delta_i} \\ \alpha_i = \alpha_1 \Delta_i + \frac{1}{V} \Phi_i, i = 2, \ldots, m \end{cases} \tag{5}$$

4.2 Last Installment Scheduling

The last installment in Fig. 1 shows the time diagram for a blocking mode communication system in which all the processors finish their tasks at the same time. Hence, we have

$$T(\beta_1) = o_1 + z_1 \beta_1 V + s_1 + w_1 \beta_1 V \tag{6}$$

$$T(\beta_i) = \sum_{j=1}^{i-1} (o_j + z_j \beta_j V) + o_i + z_i \beta_i V + s_i + w_i \beta_i V \tag{7}$$

Processing time for all processors is the same, $T(\beta_m) = T(\beta_{m+1})$, hence

$$s_i + w_i \beta_i V = o_{i+1} + z_{i+1} \beta_{i+1} V + s_{i+1} + w_{i+1} \beta_{i+1} V \tag{8}$$

$$\beta_{i+1} = \frac{s_i - (s_{i+1} + o_{i+1})}{V(w_{i+1} + z_{i+1})} + \frac{w_i}{w_{i+1} + z_{i+1}} \beta_i \tag{9}$$

Again, we define two new symbols, $\delta_{i+1} = \frac{s_i - (s_{i+1}+o_{i+1})}{w_{i+1}+z_{i+1}}$ and $\varepsilon_{i+1} = \frac{w_i}{w_{i+1}+z_{i+1}}$. Eq.(9) is rewritten as

$$\beta_{i+1} = \frac{1}{V}\delta_{i+1} + \varepsilon_{i+1}\beta_i \tag{10}$$

We assume $\delta_i \geq 0$. This assumption is true when size of job is large enough that all the processors participate in the job. After solving Eq.(10), we have $\mathrm{E}_i = \prod_{j=2}^{i} \varepsilon_j$ and $\Gamma_i = \sum_{j=2}^{i}(\delta_j \prod_{k=j+1}^{i} \varepsilon_k)$. We know that $\sum_{i=1}^{m} \beta_i = 1$. Therefore,

$$\begin{cases} \beta_i = \mathrm{E}_i\beta_1 + \frac{1}{V}\Gamma_i, i = 2, \ldots, m \\ \beta_1 = \frac{1 - \frac{1}{V}\sum_{i=2}^{m}\Gamma_i}{1+\sum_{i=2}^{m}\mathrm{E}_i} \end{cases} \tag{11}$$

By using Eq.(11), we can easily schedule a task in a heterogeneous environment with blocking mode communication which includes communication and computation overheads.

4.3 The Proper Number of Processors

The number of processors is important for a good scheduling. If the number of processors is increased, the waiting time for getting a task is increased for each processor.

As mentioned, we should decrease idle time for each processor. We should, therefore, try to have the same time for computing a task in a processor and total time taken for transferring data to the others. Since we have a Computation Based system, we should not delay computation but we can have waiting time before transferring data. Therefore, we have,

$$\alpha_1 V w_1 + s_1 \geq \sum_{j=2}^{m} \alpha_j V z_j + o_j \tag{12}$$

This means that computation time for the first processor should be larger or equal to the sum of all other processors' communication time in each installment. The best state to achieve is when they are equal because the system does not have any idle time due to computation. In this equation, the unknown parameter is m, the proper number of processors. With Eq.(5), we rewrite Eq.(12) as

$$\alpha_1 V w_1 + s_1 \geq \sum_{j=2}^{m} [(\alpha_1 \Delta_i + \frac{1}{V}\Phi_i) V z_j + o_j] \tag{13}$$

$$\alpha_1 V(w_1 - \sum_{j=2}^{m} \Delta_j z_j) \geq \sum_{j=2}^{m} (\Phi_j z_j + o_j) - s_1 \tag{14}$$

We replace α_1 with its equation in Eq.(5)

$$\begin{aligned} V(w_1 - \sum_{j=2}^{m} \Delta_j z_j) \geq & (1 + \sum_{i=2}^{m} \Delta_i) \left[\sum_{j=2}^{m} (\Phi_j z_j + o_j) - s_1 \right] \\ & + \sum_{i=2}^{m} \Phi_i (w_1 - \sum_{j=2}^{m} \Delta_j z_j) \end{aligned} \tag{15}$$

The proper number of installments is called $n+1$. Since we have stated that the size for all installments is the same, the size of task for each internal installment is $V = \frac{W}{n+1}$. Thus,

$$\begin{aligned} \frac{W}{n+1}(w_1 - \sum_{j=2}^{m} \Delta_j z_j) \geq & (1 + \sum_{i=2}^{m} \Delta_i) \left[\sum_{j=2}^{m} (\Phi_j z_j + o_j) - s_1 \right] + \\ & \sum_{i=2}^{m} \Phi_i (w_1 - \sum_{j=2}^{m} \Delta_j z_j) \end{aligned} \tag{16}$$

$$W \geq \left[\frac{(1 + \sum_{i=2}^{m} \Delta_i) \left[\sum_{j=2}^{m} (\Phi_j z_j + o_j) - s_1 \right]}{(w_1 - \sum_{j=2}^{m} \Delta_j z_j)} + \sum_{i=2}^{m} \Phi_i \right] (n+1) \tag{17}$$

From Eq.(17), we find that the proper number of processors is related to the number of installments. Therefore, before solving Eq.(17), we should calculate the proper number of installments. In the remainder of this paper, we use m to show the proper number of processors.

4.4 The Proper Number of Installments

Finding the proper number of installments is one of the main parts of scheduling heterogeneous multi-installment systems.

With reference to Fig. 1, by using Eq.(5) and Eq.(11), we arrive at Eq.(18) to find the response time of the job ($T(W)$).

$$T(W) = n(\alpha_1 V(w_1 + z_1) + o_1 + s_1) + \beta_1 V(w_1 + z_1) + o_1 + s_1 \tag{18}$$

We replace α_1 and β_1 with their values in Eq.(5) and Eq.(11) respectively.

$$\begin{aligned} T(W) =& n\frac{V-\sum_{i=2}^{m}\Phi_i}{1+\sum_{i=2}^{m}\Delta_i}(z_1+w_1)+n(o_1+s_1)+ \\ & \frac{V-\sum_{i=2}^{m-1}\Gamma_i}{1+\sum_{i=2}^{m-1}\mathrm{E}_i}(z_1+w_1)+(o_1+s_1) \end{aligned} \tag{19}$$

Therefore, with reference to $V=\frac{W}{n+1}$, we rewrite Eq.(19) as

$$\begin{aligned} T(W) =& n\frac{\frac{W}{n+1}-\sum_{i=2}^{m}\Phi_i}{1+\sum_{i=2}^{m}\Delta_i}(z_1+w_1)+n(o_1+s_1)+ \\ & \frac{\frac{W}{n+1}-\sum_{i=2}^{m-1}\Gamma_i}{1+\sum_{i=2}^{m-1}\mathrm{E}_i}(z_1+w_1)+(o_1+s_1) \end{aligned} \tag{20}$$

One of the known methods for finding the minimum or maximum value of an equation is derivation. We calculate the derivative of Eq.(20) based on n.

$$\begin{aligned} & T'(W)=\frac{-(z_1+w_1)\sum_{i=2}^{m}\Phi_i}{1+\sum_{i=2}^{m}\Delta_i}+\frac{-W(z_1+w_1)}{(n+1)^2(1+\sum_{i=2}^{m-1}\mathrm{E}_i)}+(o_1+s_1)+ \\ & \frac{W(z_1+w_1)(n+1)(1+\sum_{i=2}^{m}\Delta_i)-nW(1+\sum_{i=2}^{m}\Delta_i)(z_1+w_1)}{(n+1)^2(1+\sum_{i=2}^{m}\Delta_i)^2} \end{aligned} \tag{21}$$

Now we set the equation to equal to zero and calculate n.

$$\begin{aligned} & \frac{W(z_1+w_1)(\sum_{i=2}^{m-1}\mathrm{E}_i-\sum_{i=2}^{m}\Delta_i)}{(1+\sum_{i=2}^{m-1}\mathrm{E}_i)[(z_1+w_1)\sum_{i=2}^{m}\Phi_i-(1+\sum_{i=2}^{m}\Delta_i)(o_1+s_1)]} \\ & =(n+1)^2 \end{aligned} \tag{22}$$

Using Eq.(22), the proper number of installments can easily be found; n is the number of internal installments and 1 refers to the last installment.

Using Eq.(17) and Eq.(22), the proper number of processors can be found by

$$\begin{aligned} W \geq & \left[\frac{(1+\sum_{i=2}^{m}\Delta_i)(\sum_{j=2}^{m}\Phi_j z_j+o_j-s_1)}{(w_1-\sum_{j=2}^{m}\Delta_j z_j)}+\sum_{i=2}^{m}\Phi_i\right]^2 . \\ & \frac{(z_1+w_1)(\sum_{i=2}^{m-1}\mathrm{E}_i-\sum_{i=2}^{m}\Delta_i)}{(1+\sum_{i=2}^{m-1}\mathrm{E}_i)[(z_1+w_1)\sum_{i=2}^{m}\Phi_i-(1+\sum_{i=2}^{m}\Delta_i)(o_1+s_1)]} \end{aligned} \tag{23}$$

To find the proper number of processors, m, we should calculate the right side of the non-equation for different m, from the first to the last of available processors. When the right side of the equation is larger than total size of data (job size) for the first time, we select $m-1$ as the proper number of processors and call it m.

5 Experiment

A set of 50 processors with randomly produced attributes was used. The value of w is 20 times as large as the z value for all processors. This means that communication speed is much faster than computation speed.

5.1 Order of Processors

In heterogeneous multi-installment systems, the order of data distribution is very important because a bad ordering leads to more General Idle Time which causes response time to increase. Therefore, we repeated experiments with different ordering algorithms (increasing z_i, increasing w_i, and Hsu's ordering algorithm which is introduced in the next section). The results of these experiments can be seen in Fig. 2.

According to the graph scale, although the response time for ordering by z_i and w_i looks the same, there are some differences and ordering by z_i is better than ordering by w_i.

5.2 Evaluating the Proposed Method

We attempted to prove that the proposed formula to find the proper number of installments is true. Therefore, we manually increased and decreased the proper number of installments. The results of this experiment for four different sizes of job can be seen in Fig. 3a.

In this experiment, we both increase and decrease one to four units from the proper number of installments for each job size. Zero shows the calculated proper number of installments using our formula. It is clear that the response time, for each job, at this zero point is less than others. Thus, our formula for calculating the proper number of installments is true.

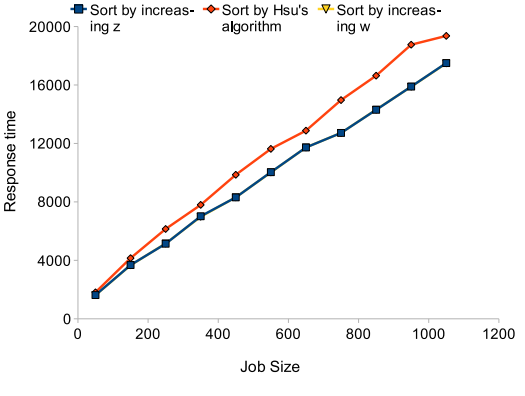
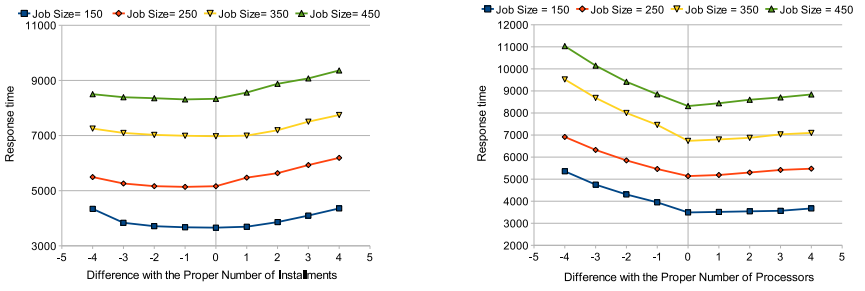


Fig. 2. Response Time vs Job Size for Different Ordering Methods



(a) **(b)**

Fig. 3. Response Time vs Difference with The Proper Number of Installments and Processors for Different Job Sizes

We did the same experiment to show that the proposed formula for finding the proper number of processors is true. The results of this experiment, presented in Fig. 3b, show that the proposed formula for calculating the proper number of processors is true.

5.3 The Proposed Methods vs Hsu's Method

We did some experiments to compare the proposed method to the presented method by Hsu et al. [10]. Hsu's method does not have any mechanism for restricting the number of used processors and it use all available processors. Therefore we did two different experiments, one experiment with all available processors (50 processors) and one experiment with the only first 16 processors,

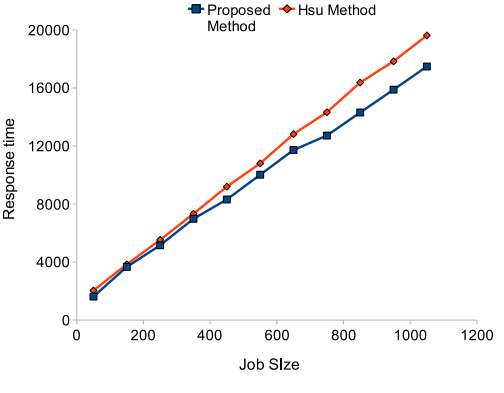


Fig. 4. Response Time vs Job Size for The Proposed Method and Hsu's Method

the calculated proper number of processors by Eq.(23), after sorting all processors with Hsu's mechanism for sorting processors. Hsu's method used a specific order of processor. They calculate $\frac{z_i}{z_i+w_i}$ for all processors and order processors by increasing this parameter.

In Fig. 4, a comparison between the proposed method for and Hsu's method can be seen. It is clear that the proposed method's response time is much better than Hsu's. One of the problems of Hsu's method is that simultaneously finishing tasks in all processors is not controlled. Another problem is the unsuitable order of processors.

6 Conclusion

Scheduling a job in a heterogeneous system which includes overheads and using multi-installment method is complex. In this paper we present a method for scheduling jobs in a multi-installment heterogeneous system which includes overheads with blocking mode communication. This method consists of four closed-form formulas for calculating the proper number of processors, the proper number of installments, chunk sizes for internal installments and chunk sizes for the last installment. We showed that ordering by decreasing communication speed. Comparing the proposed method with Hsu's method showed that response time by the proposed method is smaller than Hsu's method.

Acknowledgment

The research was partially supported by the Malaysian Ministry of Higher Education, FRGS No: 01-11-09-734FR.

References

1. Robertazzi, T.: Ten reasons to use divisible load theory. Computer 36(5), 63–68 (2003)
2. Shokripour, A., Othman, M.: Categorizing DLT researches and its applications. European Journal of Scientific Research 37(3), 496–515 (2009)
3. Mingsheng, S.: Optimal algorithm for scheduling large divisible workload on heterogeneous system. Appl. Math. Model 32, 1682–1695 (2008)
4. Mingsheng, S., Shixin, S.: Optimal multi-installments algorithm for divisible load scheduling. In: Eighth International Conference on High-Performance Computing in Asia-Pacific Region, pp. 45–54 (2005)
5. Shokripour, A., Othman, M., Ibrahim, H.: A new algorithm for divisible load scheduling with different processor available times. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010. LNCS, vol. 5990, pp. 221–230. Springer, Heidelberg (2010)
6. Shokripour, A., Othman, M., Ibrahim, H., Subramaniam, S.: A new method for job scheduling in a non-dedicated heterogeneous system. Accepted in Procedia Computer Science (2010)

7. Yang, Y., Casanova, H.: Umr: A multi-round algorithm for scheduling divisible workloads. In: 17th International Symposium on Parallel and Distributed Processing (2003)
8. Yamamoto, H., Tsuru, M., Oie, Y.: A parallel transferable uniform multi-round algorithm in heterogeneous distributed computing environment. In: Gerndt, M., Kranzlmüller, D. (eds.) HPCC 2006. LNCS, vol. 4208, pp. 51–60. Springer, Heidelberg (2006)
9. Lee, D., Ramakrishna, R.S.: Inter-round scheduling for divisible workload applications. In: Hobbs, M., Goscinski, A.M., Zhou, W. (eds.) ICA3PP 2005. LNCS, vol. 3719, pp. 225–231. Springer, Heidelberg (2005)
10. Hsu, C.H., Chen, T.L., Park, J.H.: On improving resource utilization and system throughput of master slave job scheduling in heterogeneous systems. J. Supercomput. 45(1), 129–150 (2008)

Subspace Entropy Maps for Rough Extended Framework

Dariusz Małyszko and Jarosław Stepaniuk

Department of Computer Science
Biaystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
{d.malyszko,j.stepaniuk}@pb.edu.pl

Abstract. Dynamic increase in development of data analysis techniques that has been strengthened and accompanied by recent advances witnessed during widespread development of information systems that depend upon detailed data analysis, require more sophisticated data analysis procedures and algorithms. In the last decades, deeper insight into data structure has been more many innovative data analysis approaches have been devised in order to make possible.

In the paper, in the **R**ough **E**xtended **F**ramework, SEM - a new family of the rough entropy based image descriptors has been introduced. The introduced rough entropy based image descriptors are created by means of introduced k-Subspace notion. The Subspace Entropy Maps analysis seems to present potentially robust medium during detailed data analysis. The material has been presented by examples of the introduced solutions as image descriptors.

1 Introduction

In the incoming development period in computer science the problem of integration of new data analysis techniques seems to be of high priority. In computer science, Recent advances witnessed during widespread development of information systems that depend upon detailed data analysis, require more sophisticated data analysis procedures and algorithms. In the last decades, deeper insight into data structure has been more many innovative data analysis approaches have been devised in order to make possible.

Data analysis based on the fuzzy sets depends primarily on the assumption, stating that data objects may belong in some degree not only to one concept or class but may partially participate in other classes. Rough set theory on the other hand assigns objects to class lower and upper approximations on the base of complete certainty about object belongingness to the class lower approximation and on the determination of the possible belongingness to the class upper approximation. Probabilistic approaches have been developed in several rough set settings, including decision-theoretic analysis, variable precision analysis, and information-theoretic analysis. Most often, probabilistic data interpretation depends upon rough membership functions and rough inclusion functions.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 42–51, 2011.

The theory of rough sets [7] has proved to be advantageously useful in managing uncertainty that arises from granularity in the domain of analysed data spaces. The practical effectiveness of the theory has been investigated in the areas of artificial and computational intelligence, for data representation and reasoning with imprecise knowledge, data classification and analysis.

Rough **E**xtended **F**ramework presents extensively developed method of data analysis. In the paper, a new family of k-Subspaces has been introduced. k-Subspaces appear to create a robust mathematical tool designed for implementation of the rough entropy based image descriptors. The introduced rough entropy based image descriptors are created by means of introduced k-Subspace notion. The Subspace Entropy Maps analysis seems to present potentially robust medium during detailed data analysis.

The paper has been structured in the following way. In Section 2 C-REF framework description has been given. In Section 3 distance and membership measures in C-REF framework have been presented. In Section 3 *REF*, RECA and SEM concepts have been described. In Section 4 an generalization of the SEM concept in the form of aggregate SEM has been described. The paper has been concluded by remarks on future research.

2 Rough Extended Clustering Framework - C-REF

2.1 Rough Set Theory Essentials

An information system is a pair (U, A) where U represents a non-empty finite set called the universe and A a non-empty finite set of attributes. Let $B \subseteq A$ and $X \subseteq U$. Taking into account these two sets, it is possible to approximate the set X making only the use of the information contained in B by the process of construction of the lower and upper approximations of X and further to express numerically the roughness $R(AS_B, X)$ of a set X with respect to B by assignment

$$R(AS_B, X) = 1 - \frac{card(LOW(AS_B, X))}{card(UPP(AS_B, X))}. \tag{1}$$

In this way, the value of the roughness of the set X equal 0 means that X is crisp with respect to B, and conversely if $R(AS_B, X) > 0$ then X is rough (i.e., X is vague with respect to B). Detailed information on rough set theory is provided in [8,10]. During last decades, rough set theory has been developed, examined and extended in many innovative probabilistic frameworks as presented in [11]. Variable precision rough set model VPRS improves upon rough set theory by the change of the subset operator definition, designed to analysis and recognition of statistical data patterns. In the variable precision rough set setting, the objects are allowed to be classified within an error not greater than a predefined threshold. Other probabilistic extensions include decision-theoretic framework and Bayesian rough set model.

In general **R**ough **E**xtended **F**ramework data object properties and structures are analyzed by means of their relation to the selected set of data objects from

the data space. This reference set of data objects performs as the set of thresholds or the set of cluster centers.

Rough Extended Entropy Framework in image segmentation has been primarily introduced in [6] in the domains of image thresholding routines, it means forming Rough Extended Entropy Thresholding Framework. In [6], rough set notions - lower and upper approximations have been applied into image thresholding. This thresholding method has been extended into multilevel thresholding for one-dimensional and two-dimensional domains - [3] rough entropy notion have been extended into multilevel granular rough entropy evolutionary thresholding of 1D data in the form of *2D-GMRET-GD* algorithm. Additionally, in [1] the authors extend this algorithm into *2D-GMRET-GD* thresholding routine of 2D image data. Further, rough entropy measures have been employed in image data clustering setting in [2], [4] described as **R**ough **E**ntropy **C**lustering **A**lgorithm.

In this context, **R**ough **E**xtended **F**ramework basically consists of two interrelated approaches, namely thresholding approach as **R**ough **E**xtended **T**hresholding **F**ramework - F-RET and clustering approach as **R**ough **E**xtended **C**lustering **F**ramework - C-REF. Each of these approaches gives way development and calculation of rough measures. Rough measures based upon entropy notion are further referred to as rough entropy measures. Cluster centers are regarded as representatives of the clusters. The main assumption made during *REF* and C-REF based analysis consists on the remark that the way data objects are distributed in the clusters determines internal data structure. In the process of the inspection of the data assignment patterns in different parametric settings it is possible to reveal or describe properly data properties.

2.2 General *REF* and RECA Concepts - Rough Entropy Measures

Rough (entropy) measures, considered as a measure of quality for data clustering gives possibility and theoretical background for development of robust clustering schemes. These clustering algorithms incorporate rough set theory, fuzzy set theory and entropy measure. Three basic rough properties that are applied in clustering scheme include

1. selection of the threshold metrics (crisp, fuzzy, probabilistic, fuzzified probabilistic) - tm,
2. the threshold type (thresholded or difference based) - tt,
3. the measure for lower and the upper approximations - crisp, fuzzy, probabilistic, fuzzified probabilistic - ma.

Data objects are assigned to lower and upper approximation on the base of the following criteria:

1. assignment performed on the basis of the distance to cluster centers within given threshold value,
2. assignment performed on the basis of the difference of distances to the cluster centers within given threshold value.

In general Rough Entropy Framework data object are analyzed by means of their relation to the selected number of cluster centers. Cluster centers are regarded as representatives of the clusters. The main assumption made during *REF* based analysis consists on the remark that the way data objects are distributed in the clusters determines internal data structure. In the process of the inspection of the data assignment patterns in different parametric settings it is possible to reveal or describe properly data properties. In the *REF* approach, the following properties are included during calculations

1. data objects, selected number of cluster centers
2. threshold type: difference, difference
3. threshold metric: crisp, fuzzy, probabilistic, fuzzified probabilistic
4. approximation measure: crisp, fuzzy, probabilistic, fuzzified probabilistic

In crisp setting, RECA measures are calculated on the base of the crisp metric. In rough clustering approaches, data points closest to the given cluster center or sufficiently close relative to the selected threshold type, are assigned to this cluster lower and upper approximations. The upper approximations are calculated in the specific, dependant upon threshold type and measure way presented in the subsequent paragraphs. Standard crisp distance most often applied in many working software data analysis systems depends upon Euclidean distance or Minkowsky distance, calculated as follows

$$d_{cr}(x_i, C_m) = \left(\Sigma_{j=1}^{d}(x_{ij} - C_{mj})^p\right)^{\frac{1}{p}} \tag{2}$$

Fuzzy membership value $\mu_{C_l}(x_i) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as

$$d_{fz}(x_i, C_m) = \mu_{C_l}(x_i) = \frac{d(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d(x_i, C_j)^{-2/(\mu-1)}} \tag{3}$$

where a real number $\mu > 1$ represents fuzzifier value and $d(x_i, C_l)$ denotes distance between data object x_i and cluster (center) C_l.

Probability distributions in RECA measures are required during measure calculations of probabilistic distance between data objects and cluster centers. Gauss distribution has been selected as probabilistic distance metric for data point $x_i \in U$ to cluster center C_m calculated as follows

$$d_{pr}(x_i, C_m) = (2\pi)^{-d/2}|\Sigma_m|^{-1/2} exp\left(-\frac{1}{2}(x_i - \mu_m)^T \Sigma_m^{-1}(x_i - \mu_m)\right) \tag{4}$$

where $|\Sigma_m|$ is the determinant of the covariance matrix Σ_m and the inverse covariance matrix for the C_m cluster is denoted as Σ_m^{-1}. Data dimensionality is denoted as d. In this way, for standard color RGB images $d = 3$, for gray scale images $d = 1$. Mean value for Gauss distribution of the cluster C_m has been denoted as μ_m.

In fuzzified probabilistic *RECA* measures, the probabilistic distances to all clusters are fuzzified by means of the following formulae applied to the $d_1, \ldots, d_n$ distances. Fuzzified membership value of probabilistic distance $\mu(pr)_{C_l}(x_i) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as

$$C_l(x_i) = \frac{d_{pr}(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d_{pr}(x_i, C_j)^{-2/(\mu-1)}} \tag{5}$$

2.3 RECA Component - RECA-C

The **RECA** component **RECA-C** in the clustering setting presents the set of the **RECA** cluster centers $C^{reca} = C^r$ together with the points assigned to each of the cluster.

1. the universe of discourse, U,
2. the input data, in image analysis setting, image points in image space - pixels - I,
3. the set of clusters $C^{reca} = C^r$,
4. the influence areas - crisp and fuzzy influence areas thet create data attribute partitions,
5. parameters for the RECA rough measures - $C, F, P, FP - T, D - RECA$
6. clusters and covariance matrices, new partitions, new cluster centers,
7. the set of the random cluster assignments $C(a)_i^r = \{x_0, x_1, \ldots, x_s\}$,
8. the set of the new clusters after reassignment $Re - C^{reca} = C^r$,
9. transformations $(C, F, P, FP) - (T, D)RECA^p$,

$$RECA - C = \{U, I, C^{reca}, C(a)^r\}$$

Additionally, the following blocks may be appended

$$RECA - C = \{U, I, C^{reca}, C(a)^r, \ldots, IA\}$$

3 Subspace Entropy Maps

3.1 Image Data Representation

Universe of Discourse for Image Data. As an universe of discourse U in image setting, the space of all possible data objects is taken into consideration. Image universe of discourse is represented as R^d dimensional data. Each dimension of the image data represents one of the bands or the channels of the image. Most often, contemporary data attribute resolution is equal to 8 bits. The attribute set is denoted as $A = \{A0, \ldots, AN\}$ and refers to the set of the bands or channels of the image.

Image data. In general notion, the image space may be defined as

$$I_A(A = \{A0, \ldots, AN\}) = I_0(A), I_1(A), \ldots, I_k(A)$$

In the detailed form, the image space is described as

$$I_A(A = \{A0, \ldots, AN\}) = I_0(\{A0, \ldots, AN\}),$$
$$I_1(\{A0, \ldots, AN\}), \ldots, I_k(\{A0, \ldots, AN\})$$

In case of color **RGB** images $A = \{R, G, B\}$ or $A = \{r, g, b\}$, the following image representation is possible

$$I(R, G, B) = \{I_0(r, g, b), I_1(r, g, b), \ldots, I_k(r, g, b)\}$$

or equivalently

$$I(r, g, b) = \{x_0(r, g, b), x_1(r, g, b), \ldots, x_k(r, g, b)\}$$

3.2 *k*-Subspace Definition

In the introduced k-Subspaces, as an possible input space an image space in attribute domain is considered or cluster centers of **RECA** clusters.

***k0*-Subspace Definition**
Data space after transformation of setting to 0 value of selected bit - p - of each data object.

$$b0(p, U)$$

***k1*-Subspace Definition**
Data space after transformation of seting to 1 value of selected bit - p - of each data object.

$$b1(p, U)$$

3.3 Examples of *k*-Subspaces

In this context, the following k-Subspaces are considered

1. image attribute space **IAS**,
2. **RECA** cluster centers space - **CCS**, for example **RECA-CCS**.

Image data space after transformation of setting to 0 value of selected bit - p - of each image data object.

$$I(R, G, B) = I_0(r, g, b), I_1(r, g, b), \ldots, I_k(r, g, b)$$

Cluster centers *k*-Subspace Definition
Cluster center space after transformation of seting to 0 value of selected bit - p - of each cluster center.

3.4 *REF* - SEM Algorithm

The **SEM** algorithm presents the method of generation of the entropy subspace maps. For the given image data, and **RECA** set, most often composed of **RECA** block, the following operations are performed

1. for each possible k-Subspace I_{kS}, generate the associated:
   (a) image k-Subspace,
   (b) cluster k-Subspace,
2. perform rough entropy k-Subspace $E_i = E(kS_i)$ calculations,
3. calculate Entropy Maps (on the basis of the remembered L, U).

Algorithm 1. Subspace Entropy Map calculation

for $j = 0, \ldots, k$ **do**
  foreach *Data object* $x_i = I_i$ **do**

$$Image(i,j) = b0(j, I_i(r,g,b))$$

  Determine the closest cluster C_l **for** x_i
  Increment Lower(C_l) **and Upper**(C_l) **by** 1
  foreach *Cluster* $C_m \neq C_l$ *with* $\mu_{C_m}(x_i) \geq \epsilon_{fuzz}$ **do**
    Increment Upper(C_m) **by** 1
  end
  Calculate (Subspace) Rough Entropy(i,j)

Algorithm 2. Entropy Map calculation

for $i = 0, \ldots, k$ **do**
  for $j = 0, \ldots, l$ **do**
    Calculate Subspace Entropy Maps RE(i,j)
Calculate Entropy Map $= \Sigma_{i,j} = RE(i,j)$

4 Aggregate Subspace Entropy Maps

4.1 RECA k-Subspaces Transformations

In k-Subspaces setting, it also possible to introduce the so called aggregate k-Subspaces consisted of multiple k-Subspaces, where the K set is created from: K=k1, .., kN separate k-Subspaces. In case of RECA k-Subspaces, the following operations during aggregate k-Subspace Entropy Maps are performed

1. prepare input images IMG1...N and their parameters
   (a) cluster centers CS
   (b) BD27059-RGB,0GB
   (c) RECA-S together with RECA-P, RECA parameters. RECA-P are given as (C, F, P, FP) - (D, T) - RECA for the selected RECA algorithm type

- P1 parameter set
- P2 parameter set

2. perform the following operations
   (a) calculate image k-Subspace for *k1* = 0, 1, ..., 7 (or N-bits), *k1*-images,
   (b) take image standard space *k*-image,
   (c) calculate cluster center k-Subspace for *k2* = 0, 1, ..., 7 (or N-bits), *k2*-clusters,
   (d) take standard clusters (centers) , k-clusters,
   (e) calculate RECA-S for the following sets
      i. *k*-image, *k2*-clusters, *RECA-S*(parameters),
      ii. *k1*-images, *k*-clusters, *RECA-S*(parameters),
      iii. *k*-image, *k*-clusters, *RECA-S*(parameters),
      iv. (a) *k1*, *k1*=0,1,...,7
      v. (b) *k2*, *k2*=0,1,...,7
      vi. (c) P1, P2,..., PN
3. create the following entropy maps
   (a) *k1*, *k1*=0,1,...,7, P1. Calculate the *RECA-S*(RECA-AS)(cluster assignment) in the form of power set (or approximation set), la(k), ua(k),
   (b) perform the same with the parameters
      i. all a) parameters with *P1*
      ii. all a) parameters with *P1*,...,*PN*
      iii. all b) parameters with *P1*,...,*PN*.

For the above approximations, or power sets, take them to be the lower and upper approximations. Calculate rough entropies the *RECA*(RE) for the above roughness, than calculate the entropy maps and subspace entropy maps. All calculated Subspace Entropy Maps present images *IMG1*... *N* descriptors.

4.2 Examples of RECA Entropy Maps

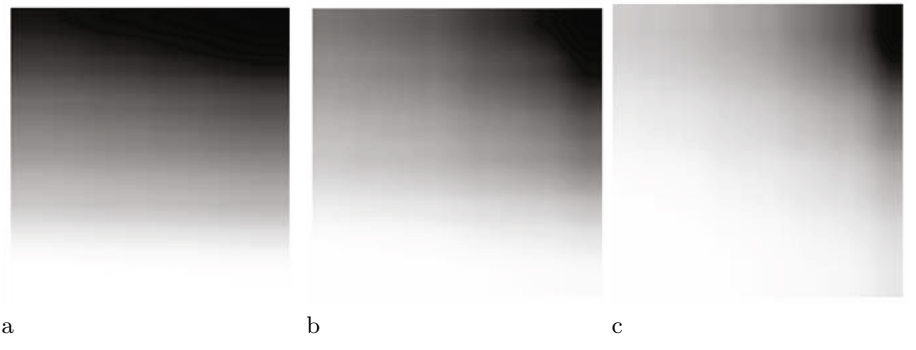


Fig. 1. Entropy maps for CFD-RECA ϵ_{fz} = 0.15-0.45, map size 20x20 - in gray-scale, (a) EM with partition 5 - 15, (b) EM with partition 10 - 10, (c) EM with partition 15 - 5

a b c

Fig. 2. Entropy maps for FFD-RECA $\epsilon_{fz} = 0.15\text{-}0.45$, map size 20x20 - in gray-scale, (a) EM with partition 5 - 15, (b) EM with partition 10 - 10, (c) EM with partition 15 - 5

5 Conclusions and Future Research

In the study, the definition, detailed analysis and presentation material of the subspace entropy maps have been presented. The combination of crisp, fuzzy, probabilistic and fuzzified probabilistic rough measures together with application of different k-Subspaces presents suitable tool for image segmentation and analysis. In this context, further research directed into extension and incorporation of the introduced k-Subspaces and rough (entropy) measures in the area of image descriptors for image analysis multimedia applications seems to be a promising area for development of modern inteligent information systems.

Acknowledgments

The research is supported by the Rector's grant of Bialystok University of Technology.

References

1. Malyszko, D., Stepaniuk, J.: Granular Multilevel Rough Entropy Thresholding in 2D Domain. In: IIS 2008: 16th International Conference Intelligent Information Systems, Zakopane, Poland, June 16-18, pp. 151–160 (2008)
2. Małyszko, D., Stepaniuk, J.: Standard and Fuzzy Rough Entropy Clustering Algorithms in Image Segmentation. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 409–418. Springer, Heidelberg (2008)
3. Malyszko, D., Stepaniuk, J.: Adaptive multilevel rough entropy evolutionary thresholding. Information Sciences 180(7), 1138–1158 (2010)
4. Malyszko, D., Stepaniuk, J.: Adaptive Rough Entropy Clustering Algorithms in Image Segmentation. Fundamenta Informaticae 98(2-3), 199–231 (2010)

5. Malyszko, D., Stepaniuk, J.: Probabilistic Rough Entropy Measures in Image Segmentation. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 40–49. Springer, Heidelberg (2010)
6. Pal, S.K., Shankar, B.U., Mitra, P.: Granular computing, rough entropy and object extraction. Pattern Recognition Letters 26(16), 2509–2517 (2005)
7. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
8. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley Sons, New York (2008)
9. Skowron, A., Stepaniuk, J.: Tolerance Approximation Spaces. Fundamenta Informaticae 27(2-3), 245–253 (1996)
10. Stepaniuk, J.: Rough–Granular Computing in Knowledge Discovery and Data Mining. Springer, Heidelberg (2008)
11. Slezak, D., Ziarko, W.: The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning 40, 81–91 (2005)

Rough Sets Applied to the RoughCast System for Steel Castings

Stanisława Kluska-Nawarecka[1,3], Dorota Wilk-Kołodziejczyk[2], Krzysztof Regulski[3], and Grzegorz Dobrowolski[3]

[1] Foundry Research Institute in Krakow, Poland
[2] Andrzej Frycz Modrzewski Krakow University, Poland
[3] AGH University of Science and Technology, Krakow, Poland
nawar@iod.krakow.pl, wilk.kolodziejczyk@gmail.com, regulski@tempus.metal.agh.edu.pl, grzela@agh.edu.pl

Abstract. Rough logic and rough sets theory are mainly devoted to analysis of incomplete, uncertain, and inconsistent data. An exploited in the article result of the theory is the so-called rough information system that gives a model for dealing with the situation in which some real objects are represented ambiguously via defined approximation (alternatives of values of the objects attributes). The proposed rough system opens a way to process such information and moreover is equipped with a powerful query language and possibility to classify real objects based on their stored rough characteristics. The implemented RoughCast system and, especially prepared for the case, knowledge base for steel casting defects form such the rough information system. They can be successfully applied in the process of knowledge integration concerning production of steel castings and, in consequence, as a tool in solving technological problems in operating foundries.

Keywords: uncertain knowledge representation, rough logic (sets), rough information systems, classification, casting defects, steel castings, computerized supporting tools.

1 Introduction

Research carried out at Foundry Research Institute in Cracow in collaboration with team from Faculty of Metals Engineering and Industrial Computer Science at AGH University of Science and Technology in Cracow on development of Information Systems for Foundry Industry [1] embraces exploration of a wide spectrum of knowledge representation methods in order to evaluate their usability and capability to be a core for specialized software tools in the field.

Although rough sets theory was invented over 20 years ago [2], it has not cause such avalanche of application attempts and real applications as e.g. fuzzy sets theory or artificial neural networks. Nevertheless, it can be noticed that the rough sets create a sound base for acquisition, processing and interpretation of incomplete, uncertain, and inconsistent knowledge. So-called engineering knowledge, dealing with engineering practice especially, has such characteristic very often. Rough sets theory [2] opens

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 52–61, 2011.

perspective for employing rough - in the sense characterized above - technological data into decision-making and classifying systems.

As a pilot implementation, RoughCast system has been built, which is a tool for classification of casting defects basing on the values of their observed attributes. The observation is confronted by the system with general and thus rough knowledge about the defects stored in the rough information system under consideration. In the reported case the knowledge about cast steel defects extracted from various standards published in different countries (Polish, Czech, French) is applied.

The additional goal of the report work is to prove that The Pawlak's rough sets theory can be applied in industry information systems and its application produces new capabilities in processing and interpretation of rough knowledge.

The article is organized as follows. The next section provides a reader with some definitions and ideas of rough sets theory necessary to show (in section 3) how knowledge included in the standards can be translated into a rough information system. As such system is armed with a powerful query language, the knowledge structuralized this way can be flexible viewed, what can produce semi-products for various kinds of the experts' activities. Section 4 presents some elements of the prototype RoughCast which is built to do the thing. The whole is illustrated with examples how the queries are formulated, calculated and evaluated (query results are obviously rough and the "roughness" is calculated also).

2 Query Language in Information Systems

Information systems use the language of mathematical logic in describing the reality to which they relate. In this respect, the rough logic uses the same semantics as the first-order logic. Largely based on the language of the set theory, in description of reality it uses terms, i.e. formal expressions, which consist of:

- descriptors, i.e. pairs (a_j, v_{ij}) ascribing certain value to the j^{th} attribute of object x_i;
- constants 0,1 (false, true);
- symbols of operators of the upper and lower approximation $\overline{S}$, $\underline{S}$;
- symbols of logical operators: negation, conjunction, alternative, implication and equivalence ($\neg$, $\cdot$, $+$, $\rightarrow$, $\leftrightarrow$).

The query language in information systems consists of a set of rules determining the construction of queries to enable the extraction of information contained in the system. Queries may assume the following forms:

- intersectional – the result is the set of items that meet the predetermined conditions;
- digital – the result is the number of elements in the set of responses;
- relational – the result is the set of objects that remain in a preset relationship with each other;
- numerical – the result of a query is the result of a computational task done on attributes;
- logical – the result may be either *true* or *false*.

If the system represents information which is a generalization of the database in which each tuple is the embodiment of a relationship regarded as a subset of the Cartesian product (data patterns or templates), the semantics of each record is defined with a logical formula [6]:

$$\varphi_i = [A_1 = v_{i,1}] \wedge [A_2 = v_{i,2}] \wedge \ldots \wedge [A_n = v_{i,n}] \tag{1}$$

The notation $A_j = v_{i,j}$ means that the formula $_i$ is true for all values that belong to the set $v_{i,j}$. Hence, if $v_{i,j} = \{v_1, v_2, v_3\}$, $A_j = v_{pj}$ means that $A_i = v_1 \vee A_i = v_2 \vee A_i = v_3$, while the array has a corresponding counterpart in the formula:

$$\Psi = \varphi 1 \vee \varphi 2 \vee .. \vee \varphi m \tag{2}$$

If the array represents some rules, the semantics of each line is defined as a formula:

$$\rho_i = [A_1 = v_{i,1}] \wedge [A_2 = v_{i,2}] \wedge \ldots \wedge [A_n = v_{i,n}] \Rightarrow [H = h_i] \tag{3}$$

On the other hand, to the array of rules is corresponding a conjunction of formulas describing the lines.

3 Knowledge of Casting Defects as an Information System

The research carried out by a Knowledge Engineering Group in the Section of Computer Science at the Faculty of Metals Engineering and Industrial Computer Science helped develop a methodology applied in the formation of decision-making tables for the knowledge of casting defects [3,4,7]. Using the developed methodology, a decision table was created for selected defects in steel castings. A fragment of the array is shown in Table 1.

Table 1. Fragment of decision table for defects in steel castings

Attribute symbol	a1	a2	a3	a4	a6	a7	a8	a9	a12
Symbol of object in array	**Damage symbol**	**Damage name**	**Standard**	**damage type**	**distribution**	**location**	**occurrence**	**damage shape**	**technological operation**
x1	341	**COLD LAPS**	CZ	wrinkles, scratch, erosion scab	local	insert wall, chaplet, surface	numerous	narrow, rounded edges	casting design, pouring, cooling
x2	W207	**COLD LAP**	PL	fissure, scratch	local	surface	single	narrow, rounded edges	gating system design, pouring
x3	W407	**COLD SHOTS**	PL	metal beads		interior		spherical	gating system design, pouring
x4	C311	**COLD LAP, COLD SHOTS**	FR	discontinuity, fissure	widespread	surface, subsurface area	numerous	rounded edges, narrow	feeding system, design, pouring
x5	C331	**COLD LAP NEAR CORE OR OTHER METALLIC PART**	FR	discontinuity	local	near inserts	data not available	curved walls	pouring, solidification

The decision table comprises a set of conditional attributes $C = \{a_4, a_5, a_6, a_7, a_8, a_9\}$ and a set of decision attributes $D = \{a_1, a_2, a_3\}$. Their sum forms a complete set of attributes $A = C \cup D$. Applying the rough set theory, it is possible to determine the elementary sets in this table. For example, for attribute a_4 (*damage type*), the elementary sets will assume the form of:

- $E_{\text{wrinkles}} = \{\emptyset\}$; $E_{\text{scratch}} = \{\emptyset\}$; $E_{\text{erosion scab}} = \{\emptyset\}$; $E_{\text{fissure}} = \{\emptyset\}$;
- $E_{\text{wrinkles, scratch, erosion scab}} = \{x_1\}$;

- $E_{\text{cold shots}} = \{x_3\}$;
- $E_{\text{fissure, scratch}} = \{x_2\}$;
- $E_{\text{discontinuity}} = \{ x_5 \}$;
- $E_{\text{discontinuity, fissure}} = \{ x_4 \}$;
- $E_{\text{wrinkles, scratch, erosion scab, fissure}} = \{\emptyset\}$; $E_{\text{wrinkles, scratch, erosion scab, fissure, cold shots}} = \{\emptyset\}$;
- $E_{\text{wrinkles, scratch, erosion scab, cold shots}} = \{\emptyset\}$; $E_{\text{discontinuity, fissure, cold shots}} = \{\emptyset\}$;
- $E_{\text{discontinuity, fissure, wrinkles, scratch, erosion scab}} = \{\emptyset\}$;
- etc.

Thus determined sets represent a partition of the universe done in respect of the relationship of indistinguishability for an attribute "*distribution*". This example shows one of the steps in the mechanism of reasoning with application of approximate logic. Further step is determination of the upper and lower approximation in the form of a pair of precise sets. Abstract class is the smallest unit in the calculation of rough sets. Depending on the query, the upper and lower approximation is calculated by summing up the appropriate elementary sets.

Example: Let us determine an approximation for the query (set): X= *{discontinuity, fissure}*. The lower approximation for the set X is the sum of all elementary sets, the descriptors of which assume the form of (a, X') where $X' \subseteq X$. The lower approximation for the set X= *{discontinuity, fissure}* is the sum of all elementary sets, the descriptors of which contain all subsets of the set X:

- descriptor: (*damage type, {discontinuity, fissure}*) is corresponding to an elementary set $E_{\text{discontinuity, fissure}}$
- descriptor: (*damage type, {discontinuity}*) is corresponding to an elementary set $E_{\text{discontinuity}}$
- descriptor: (*damage type,{fissure}*) is corresponding to an elementary set E_{fissure}

The sum of the above sets forms the lower approximation:

$$\underline{S}\ (\textit{damage type, X}) = E_{\text{discontinuity, fissure}} \cup E_{\text{discontinuity}} \cup E_{\text{fissure}} \tag{4}$$

$$\underline{S}\ (\textit{damage type, X}) = \{ x_4, x_5\}$$

The upper approximation for the set X is the sum of all elementary sets, the descriptors of which assume the form of (a, X') where $X' \cap X \neq \emptyset$:

- descriptor: (*damage type, {discontinuity, fissure}*) is corresponding to an elementary set $E_{\text{discontinuity, fissure}}$
- descriptor: (*damage type, {discontinuity}*) is corresponding to an elementary set $E_{\text{discontinuity}}$
- descriptor: (*damage type,{fissure}*) is corresponding to an elementary set E_{fissure}
- descriptor: (*damage type, {discontinuity, fissure}*) is corresponding to an elementary set $E_{\text{discontinuity, fissure}}$
- descriptor: (*damage type, {discontinuity}*) is corresponding to an elementary set $E_{\text{discontinuity}}$
- descriptor: (*damage type,{fissure}*) is corresponding to an elementary set E_{fissure}

- descriptor: (*damage type, {fissure, scratch}*) is corresponding to an elementary set $E_{\text{fissure, scratch}}$
- descriptor: (*damage type,{ fissure, cold shots}*) is corresponding to an elementary set $E_{\text{fissure, cold shots}}$
- etc.

The sum of the above sets forms the upper approximation:

$$\overline{S}\,(\textit{damage type, X}) = E_{\textit{discontinuity, fissure}} \cup E_{\textit{discontinuity}} \cup E_{\textit{fissure}} \cup E_{\textit{fissure,scratch}} \cup E_{\textit{fissure, cold shots}} \cup E_{\textit{discontinuity, cold shots}} \cup E_{\textit{discontinuity, scratch ...}} \quad (5)$$

$$\overline{S}\,(\textit{damage type, X}) = \{x_2, x_4, x_5\}$$

For a set of attributes of size larger than 1, the calculation of upper and lower approximations can be represented in the following way: two families of elementary sets for attributes *damage type* and *distribution* are chosen by analogy to the above example. These are for the attribute *damage type*:

- $E_{\text{discontinuity, fissure}} = \{x_4\}$
- $E_{\text{discontinuity}} = \{x_5\}$
- $E_{\text{fissure}} = \{ \emptyset \}$;

for the attribute *distribution*:

- $E_{\text{local}} = \{x_1, x_2, x_5\}$
- $E_{\text{widespread}} = \{x_4\}$
- $E_{\text{local, widespread}} = \{ \emptyset \}$;

The third family of elementary sets is obtained for the set of attributes B = (*damage type, distribution*), resulting from the Cartesian product of both attributes. To simplify the computational complexity, only non-empty, elementary sets are considered for the attribute *damage type*:

- $E_{\text{discontinuity, fissure}} = \{x_4\}$
- $E_{\text{discontinuity}} = \{x_5\}$

for the attribute *distribution*:

- $E_{\text{local}} = \{x_1, x_2, x_5\}$
- $E_{\text{widespread}} = \{x_4\}$

The elementary sets for a set of attributes B are formed in the following way:

- $E_{\text{discontinuity, local}} = E_{\text{discontinuity}} \cap E_{\text{local}} = \{x_5\}$
- $E_{\text{discontinuity, widespread}} = E_{\text{discontinuity}} \cap E_{\text{widespread}} = \{ \emptyset \}$
- $E_{\text{discontinuity, fissure, local}} = E_{\text{discontinuity, fissure}} \cap E_{\text{local}} = \{ \emptyset \}$
- $E_{\text{discontinuity, fissure, widespread}} = E_{\text{discontinuity, fissure}} \cap E_{\text{widespread}} = \{ x_4 \}$

Obtained from the Cartesian product sets can be reduced to the existing elementary sets for B.

Query example: t_1= *(damage type, {discontinuity, fissure})* · *(distribution, {local})*

When calculating the lower approximation, it is necessary to sum up all the elementary sets for the sets of attribute values which form possible subsets of sets in a query:

$\underline{S}(t_1)$ = *(damage type, {discontinuity, fissure}) · (distribution, {local}) + (damage type, {discontinuity})] · (distribution, { local})*

The result is a sum of elementary sets forming the lower approximation:

$E_{\text{discontinuity, local}} \cup E_{\text{discontinuity, fissure, local}} = \{x_5\}$

The upper approximation for the submitted query is:

$\overline{S}(t_1)$ = (*damage type*, {*discontinuity, fissure*}) · (*distribution*, {*local*}) + (*damage type*, {*discontinuity*})] · (*distribution*, {*local*}) + (*damage type*, {*discontinuity*, *scratch*})] · (*distribution*, {*local*})

The result is a sum of elementary sets forming the upper approximation:

$E_{\text{discontinuity,local}} \cup E_{\text{fissure,scratch,local}} \cup E_{\text{discontinuity,fissure,local}} = \{ x_2, x_5, \}$

Where the accuracy of approximations is:

$$\mu(t_1) = \frac{card\,\underline{S}}{card\,\overline{S}} = \frac{1}{2} \tag{6}$$

4 Implementation of Reasoning in Rough Logic for the Diagnosis of Defects in Castings - A RoughCast System

Within the application of Pawlak rough set theory, a pilot implementation in the form of a diagnostic *RoughCast* system was performed [5]. The system is web-based application accessible by the browser, which offers the functionality of an expert system, in particular the ability to classify objects based on their attributes.

The system operates on decision-making tables - this data structure enables the use of inference engine based on rough logic. The system maintains a dialogue with the user asking questions about the successive attributes. The user can select the answer (the required attribute) in an intuitive form. Owing to this system of query formation, the user does not need to know the syntax and semantics of the queries in rough logic. However, to make such dialogue possible, without the need for a user to build a query in the language of logic, the system was equipped with an interpreter of queries in a semantics much more narrow than the original Pawlak semantics. It has been assumed that the most convenient way to build queries in the case of casting defects is a selection from the list of attributes required for a given object (defect). User selects what features (attributes) the defect has. This approach is consistent with cases of the daily use when user has to deal with specific cases of defects, and not with hypothetical tuples. Thus set up queries are limited to conjunctions of attributes, and for this reason the query interpreter has been equipped with only this logical operator.

4.1 Knowledge Base for the *RoughCast* System

The knowledge base created for the *RoughCast* system is an embodiment of Pawlak information system in the form of a decision-making table. It contains tabulated knowledge on the characteristics of defects in steel castings taken from the Polish Standards, Czech studies, French catalogue and German manual of casting defects. A fragment of the decision-making table for cast steel is shown in Table 2.

In its original layout of a spreadsheet edition, the decision-making table for cast steel assumes the following form:

Table 2. Fragment of decision-making table for cast steel

	A	B	C	D	E	F
1	Defect name	Standard	Symbol	damage type	Visibility	damage size
20	Cold laps	CZ	341	wrinkles	sharply outlined	distinct
25	Crush, Bruising, Push up	CZ	116	cavity	sharply outlined	distinct
30	Dent	CZ	123	dent	sharply outlined	distinct
45	Internal contraction crack	CZ	313	discontinuity	invisible	distinct
48	Mechanical damage	PL	W101	dent	sharply outlined	distinct
53	Misrun	PL	W102	part of casting missing	sharply outlined	distinct
58	Knob	PL	W103	deformation	sharply outlined	distinct
176	Cold crack	D	21	discontinuity	sharply outlined	
179	Crack in core	D	23	buildup		
180	Hot crack	D	48	fissures	visible with naked eye	scattered
184	Mould crack	D	11	buildups	visible with naked eye	local
185				knobs		
186	Cold fracture or cold crack	FR	C111	discontinuity	hardly visible	distinct
194	Cold crack	FR	C211	fissure	sharply outlined	distinct
195	Crack	FR	C221	fissure	visible	distinct
199	Hot crack	FR	C222	fissure	sharply outlined	distinct
201	Cold lap, cold shots	FR	C311	discontinuity	sharply outlined	distinct

The RoughCast system enables the exchange of knowledge bases. While working with the system, user has the option to download the current knowledge base in the form of a spreadsheet, edit this base locally on his terminal, and update it in the system.

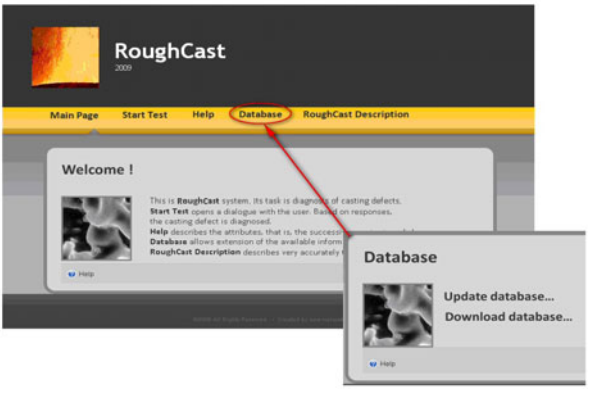


Fig. 1. Updating the knowledge base in RoughCast

The way the dialogue is carried out is directly dependent on the structure of decision table and, consequently, the system allows reasoning using tables with any knowledge, not only the one related with casting.

4.2 Dialogue with User and Reasoning in the RoughCast System

Dialogue with user in the RoughCast system starts with a form containing the individual limit values for the first conditional attribute from the decision-making table. User approves the selected attributes, owing to which the system is able to calculate the upper and lower approximation for thus formed query (fig. 1).

Based on user selection, for the successive attributes, the system creates queries as a conjunction of attributes, such as:

$t1 = (\text{damage type}, \{\text{discontinuity, fissure}\}) \cdot (\text{distribution}, \{\text{local}\})$

The computed approximations are presented to the user each time an attribute is preset, due to which the user can interrupt the dialogue any time he considers that the result satisfies him.

The system in successive steps calculates the dependency relationship for the attribute currently analysed and for the next attribute in the decision-making table. In this way, the system limits the number of questions the user is being asked, and checks whether in the selection of the next attribute to a subset of attributes, the same division of objects into decision classes is preserved as in the case of the currently examined attribute. If the partition of the universe is equivalent, the next attribute can be reduced thereby bypassing the question.

In this way, for better computational efficiency, the system of reducts known from Pawlak's theory has been simplified. Such simplification is possible because of the way in which the decision-making table is prepared. The dialogue is carried out until

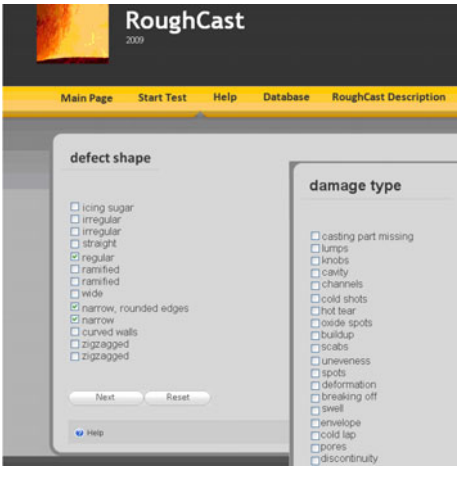


Fig. 2. Forms selecting attribute values in the RoughCast system

all the queries (conditional attributes) have been used, or until the user finishes his dialogue, feeling that the result he has obtained with approximations is already satisfactory. At each stage of the reasoning, user can check the importance of each attribute referring to a "help" manual implemented in the system.

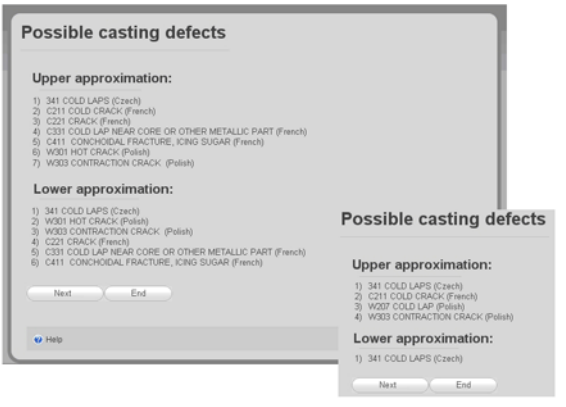


Fig. 3. Upper and lower approximations calculated in a single step of reasoning and the final result of dialogue for the example of "cold lap" defect according to the Czech classification system

The system was tested in experiments [5], which have proved that:

- final result in the form of upper and lower approximation contains the searched defect, what was confirmed by experts of the domain,
- large part of the queries was reduced and the number of attributes was limited for each query, what assures usability of the system,
- the resulting set of lower approximations contains only the searched defect, what was a main indicator of correct calculation in system tests.

5 Summary

An attempt was made to demonstrate that the rough logic as a formalism of knowledge representation, enabling solving the problem of uncertainty and incomplete information, is a convenient tool, especially in classification tasks.

The use of this formalism allows solving a number of difficulties arising from the granularity of knowledge about castings in the form of indistinguishable descriptions created with attributes and inconsistent classifications from different sources (as in the case of standards for steel castings).

The implemented RoughCast system as well as the knowledge base about steel casting defects prove that the adopted assumptions are correct, and Pawlak rough set theory can be successfully applied in the process of knowledge integration regarding the manufacture of steel castings.

References

1. Dobrowolski, G., Marcjan, R., Nawarecki, E., Kluska-Nawarecka, S., Dziadus, J.: Development of INFOCAST: Information system for foundry industry. TASK Quarterly 7(2), 283–289 (2003)
2. Pawlak, Z.: Rough sets. Int. J. of Inf. and Comp. Sci. 11(341) (1982)
3. Wilk-Kołodziejczyk D.: The structure of algorithms and knowledge modules for the diagnosis of defects in metal objects, Doctor's Thesis, AGH, Kraków (2009) (in Polish)
4. Kluska-Nawarecka, S., Wilk-Kołodziejczyk, D., Dobrowolski, G., Nawarecki, E.: Structuralization of knowledge about casting defects diagnosis based on rough set theory. Computer Methods In Materials Science 9(2) (2009)
5. Walewska E.: Application of rough set theory in diagnosis of casting defects, MSc. Thesis, WEAIIE AGH, Kraków (2010) (in Polish)
6. Ligęza, A., Szpyrka, M., Klimek, R., Szmuc, T.: Verification of selected qualitative properties of array systems with the knowledge base. In: Bubnicki, Z., Grzech, A. (eds.) Knowledge Engineering and Expert Systems. T. 1, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, pp. 103–110 (2000) (in Polish)
7. Kluska-Nawarecka, S., Górny, Z., Pysz, S., Regulski, K.: An accessible through network, adapted to new technologies, expert support system for foundry processes, operating in the field of diagnosis and decision-making. In: Sobczak, J. (ed.) Innovations in foundry, Part 3, Instytut Odlewnictwa, Kraków, pp. 249–261 (2009) (in Polish)

RECA Components in Rough Extended Clustering Framework

Dariusz Małyszko and Jarosław Stepaniuk

Department of Computer Science
Biaystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
{d.malyszko,j.stepaniuk}@pb.edu.pl

Abstract. **R**ough **E**xtended **F**ramework - REF - presents recently devised algorithmic approach to data analysis based upon inspection of the data object relation to predefined number of clusters or thresholds areas. Clusters most often are represented by the cluster center, and the cluster centers are viewed as cluster representitives. In the paper, in the **R**ough **E**xtended **C**lustering **F**ramework, the basic RECA (**R**ough **E**ntropy **C**lustering **A**lgorithms) construction blocks or components have been introduced and presented on illustrative examples. The introduced RECA components create starting point into data analysis performed on the *REF* and *C-REF* framework.

1 Introduction

In incoming years the role and impact of robust and high-quality data analysis techniques seems to become prominent. In this context, research subjects focusing on the development, performance, optimalization and speeding up of data analysis tools become one of the more important topics. In order to make data analysis methods more robust and adequate for incorporation into modern information systems development of more sophisticated systems is required. Modern data processing and analysis requires development of novel methodologies, algorithms and approaches. Furthermore, in addition to the introduction of new techniques and algorithms, most often in order to obtain better performance and it is required to combine different techniques into one coherent and more robust system.

In general **R**ough **E**xtended **F**ramework data object properties and structures are analyzed by means of their relation to the selected set of data objects from the data space [1], [3]. This reference set of data objects performs as the set of thresholds or the set of cluster centers. In the process of the inspection of the data assignment patterns in different parametric settings it is possible to reveal or describe properly data properties. In this rough framework an assumption is made that metric and other relation between data objects and cluster (most often represented by cluster centers) is informative of the internal data structure. On the other hand, by inspecting internal data structure, detailed and precise

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 62–71, 2011.

insight into data object relations, exploration of the dependencies is made more robust.

In the present research state of the art, main *REF* domains are - *T-REF* - rough analysis based upon examining of the data object assignment and distance relation relative to thresholds, see [4] for details. The other *REF* domain consists of *C-REF* - rough sets based data analysis that performs insight into data structure by examining data metric relation relative to selected number of clusters. RECA analysis depends upon interpretation of the data object distance relation to the selected set of the clusters. Most often as a distance of the cluster to the data object is the distance of the object to the cluster center. The main area of application of *REF* based theories and methodologies is targeted into

- data classification algorithms,
- image data segmentation algorithms,
- robust image descriptors,
- uncertainty handling by means of incorporation of rough, fuzzy and probabilistic notions into integrated and uniform data analysis setting.

This paper has been structured in the following way. The basic notions of *C-REF* Framework have been presented in Section 2. In Section 3 the information about distance and membership measures in C-REF Framework has been presented. In Section 4 the idea and notions of the *RECA* components have been described. The operations on C-REF components have been described in Section 5 followed by concluding remarks.

2 Rough Extended Clustering Framework Concepts

In *C-REF* framework, data analysis is performed on the inspection of the data objects to the predefined number of clusters. Each cluster has assigned its data centers that represent this cluster. The data objects are assigned to the cluster on the base of some selected criteria mainly based on the distance to the closest clusters. The distance of the data point to the cluster most often is calculated on the basis of the distance of the data point to the closest cluster center or cluster centers. Most often this kind of procedures requires selection of the proper distance metric and further on the assumption of the distance type, it means crisp distance, fuzzy distance, probabilistic distance, fuzzy probabilistic distance. This clustering based data analysis procedures require the definition of the distance between data objects and cluster centers. In the *REF* framework, a new family of distance measures in the form of the hypersphere d-metric has been introduced. The hypersphere d-metric has been introduced and presented in [2]. The hypersphere d-metric, in case of taking standard distance and Voronoi distance performs as standard distance during k-means based clustering routines.

RECA sets are devised in order to make more advantageous benefits from combining of rough, fuzzy and probabilistic approaches in clustering methodology of data analysis. *RECA* sets are primarily introduced in *REF* framework that aims at making data analysis on the basis of rough approach. *RECA* part of

REF framework deals with application of clustering techniques into data analysis algorithms. In the subsequent material, the following definition of the *RECA* cluster component has been proposed and introduced

Table 1. RECA-Components

The term	Description
RECA-C	clusters, especially with selected *RECA-CC* cluster centers
RECA-P	parameters
RECA-AS	assignment strategy to clusters
RECA-IA	area, region determined by the set of *RECA-AS*
RECA-R	area, region determined by the set of *RECA-AS*
RECA-S	*RECA-C*, *RECA-P* and *RECA-A*
RECA-B	combination of predefined number of RECA-S sets
	referred to as RECA-B blocks, especially RECA-N-Blocks
RECA-ST	understood as operation on *RECA-S* or *RECA-B*

Rough measures, considered as a measure of quality for data clustering gives possibility and theoretical background for development of robust clustering schemes. These clustering algorithms incorporate rough set theory, fuzzy set theory and entropy measure. The three basic rough operations that are applied during rough clustering based data analysis scheme include the following operations

1. Assignment to the RECA-S Set by the means of
   (a) selection of the threshold metrics (crisp, fuzzy, probabilistic, fuzzified probabilistic) - tm,
   (b) the threshold type (thresholded or difference based) - tt,
   (c) Data objects are assigned to *RECA-S* (lower and upper approximations) on the base of the following criteria (see Table 2)
      i. assignment performed on the basis of the distance to cluster centers within given threshold value,
      ii. assignment performed on the basis of the difference of distances to the cluster centers within the given threshold value.
2. Calculation of the RECA-S measure by the means of selection of the following membership measures (see Table 3 for detailed information)
   (a) crisp,
   (b) fuzzy,
   (c) probabilistic,
   (d) fuzzified probabilistic,
3. Operations on RECA-S Sets (described in Section 5).

3 The Distance and Membership Measures in C-REF Framework

3.1 Generalized Hypersphere *d*-Metric in C-REF Framework

In the paper, the definition of the notion of the generalized hypersphere d-metric - or cluster in standard setting - has been accommodated as presented in [2].

Let $\oplus = \{\phi_1, \ldots, \phi_n\}$ denotes a set of so-called generalized (Mobius) sites of R^d with ϕ_i meaning

$$C_i^h = \phi_i = (p_i, \lambda_i, \mu_i, r, s, M_i) \tag{1}$$

that is formed of a point p_i of R^d, and four real numbers λ_i, μ_i, r and s. Additionally, M represents matrix. For a point $x \in R^d$, the distance $\delta_i(x, \phi_i)$ from x to the generalized (Mobius) site ϕ_i is defined as

$$\delta_i(x, \phi_i) = \delta_i(x, C_i^h) = \lambda_i \left[(x - p_i)^r\right]^T M_i (x - p_i)^s - \mu_i \tag{2}$$

For $r = s = 1.0$ the formulae is

$$\delta_i(x, \phi_i) = \delta_i(x, C_i^h) = \lambda_i \left(x - p_i\right)^T M_i (x - p_i) - \mu_i \tag{3}$$

The data point x is the closest to the Mobius site δ_i that the distance $\delta_i(x, \phi_i)$ is the lowest.

3.2 *RECA-S* Membership Measures in *C-REF* Framework

Crisp RECA Measures - In crisp setting, *RECA* measures are calculated on the base of the crisp metric. In rough clustering approaches, data points closest to the given cluster center or sufficiently close relative to the selected threshold type, are assigned to this cluster lower and upper approximations. The upper approximations are calculated in the specific, dependant upon threshold type and measure way presented in the subsequent paragraphs. Standard crisp distance most often applied in many working software data analysis systems depends upon Euclidean distance (p=2) or Minkowsky distance (p ¿ 0), calculated as follows

$$d_{cr}(x_i, C_m) = \left(\Sigma_{j=1}^{d} (x_{ij} - C_{mj})^p\right)^{\frac{1}{p}} \tag{4}$$

Fuzzy RECA Measures - Fuzzy membership value $\mu_{C_l}(x_i) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as

$$d_{fz}(x_i, C_m) = \mu_{C_l}(x_i) = \frac{d(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d(x_i, C_j)^{-2/(\mu-1)}} \tag{5}$$

where a real number $\mu > 1$ represents fuzzifier value and $d(x_i, C_l)$ denotes distance between data object x_i and cluster (center) C_l.

Probabilistic RECA Measures - Probability distributions in *RECA* measures are required during measure calculations of probabilistic distance between

data objects and cluster centers. Gauss distribution has been selected as probabilistic distance metric for data point $x_i \in U$ to cluster center C_m calculated as follows

$$d_{pr}(x_i, C_m) = (2\pi)^{-d/2}|\Sigma_m|^{-1/2} exp\left(-\frac{1}{2}(x_i - \mu_m)^T \Sigma_m^{-1}(x_i - \mu_m)\right) \quad (6)$$

where $|\Sigma_m|$ is the determinant of the covariance matrix Σ_m and the inverse covariance matrix for the C_m cluster is denoted as Σ_m^{-1}. Data dimensionality is denoted as d. The mean value for Gauss distribution of the cluster C_m has been denoted as μ_m.

Fuzzified Probabilistic RECA Measures - In fuzzified probabilistic *RECA* measures, the probabilistic distances to all clusters are fuzzified by means of the following formulae applied to the $d_1, \ldots, d_n$ distances. Fuzzified membership value of probabilistic distance $\mu(pr)_{C_l}(x_i) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as

$$C_l(x_i) = \frac{d_{pr}(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d_{pr}(x_i, C_j)^{-2/(\mu-1)}} \quad (7)$$

Table 2. Threshold distance metric and Difference metric

Difference metric	Symbol	Condition
CD	$d_{cr}(x_i, C_m)$	$\lvert d_{cr}(x_i, C_m) - d_{cr}(x_i, C_l) \leq \epsilon_{cr}$
FD	$d_{fz}(x_i, C_m)$	$\lvert(\mu_{C_m}(x_i) - \mu_{C_l}(x_i)\rvert \leq \epsilon_{fz}$
PD	$d_{pr}(x_i, C_m)$	$\lvert d_{pr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{pr}$
FPD	$d_{fp}(x_i, C_m)$	$\lvert d_{fp}(x_i, C_m) - d_{fp}(x_i, C_l)\rvert \leq \epsilon_{fp}$
Distance metric	**Symbol**	**Condition**
CT	$d_{cr}(x_i, C_m)$	$d_{cr}(x_i, C_m) \leq \epsilon_{cr}$
FT	$d_{fz}(x_i, C_m)$	$(\mu_{C_m}(x_i) \geq \epsilon_{fz}$
PT	$d_{pr}(x_i, C_m)$	$d_{pr}(x_i, C_m) \geq \epsilon_{pr}$
FPT	$d_{fp}(x_i, C_m)$	$d_{fp}(x_i, C_m) \geq \epsilon_{fp}$

4 Components in C-REF Framework

In the subsequent material, the following definition of the *RECA* cluster component has been proposed and introduced. Additionally, *RECA* components, most often are defined and performed with the following sets are considered as *RECA* discourse and *RECA* image space. The detailed description and information about the above-mentioned *RECA* components has been presented in the subsequent sections.

Table 3. The measures for RECA-S sets (for example for RECA-S sets that represent the lower and upper approximations)

Approximation	Distance	Value
Cr	$m_{cr}(x_i, C_m)$	1
Fz	$m_{fz}(x_i, C_m)$	μ_{C_m}
Pr	$m_{pr}(x_i, C_m)$	$\|d_{pr}(x_i, C_m)$
FP	$m_{fp}(x_i, C_m)$	$\|d_{fpr}(x_i, C_m)$

4.1 RECA-C Clusters

The *RECA* cluster *RECA-C* presents the set of the *RECA* cluster centers in the form

$$C_i^h = \phi_i = (p_i, \lambda_i, \mu_i, r, s, M_i) \tag{8}$$

The *REF* or RECA set of clusters presents the set of the *RECA* cluster referred to as

$$C^{reca} = C^r = \{C_0^r, C_1^r, \ldots, C_n^r\} \tag{9}$$

where n denotes number of clusters. In this way, each hypersphere cluster center C_i^r is represented as

$$C_i^h = \phi_i = (p_i, \lambda_i, \mu_i, r, s, M_i) \tag{10}$$

In case of standard clusters, with Voronoi distance, the cluster is referred to as

$$C_i^r = \phi_i = (p_i, \lambda_i, \mu_i, r, s, M_i) \tag{11}$$

Data points assigned to the cluster C_i^r are denoted as $C(a)_i^r = \{x_0, x_1, \ldots, x_s\}$.

The parameters denoted as *RECA-P* determine the component properties such as *RECA-S* set properties.

4.2 Influence Areas - RECA-IA

Definition - influence areas describe data objects or attribute values that are assigned to the given clusters. Taking into account the fact that clusters always should be considered as the one entity with the following factors

1. crisp and fuzzy influence areas - Cr-IA, Fz-IA,
2. probabilistic influence areas - Pr-IA, P-IA,
3. fuzzified probabilistic influence areas - FzPr-IA, FP-IA.

Both influence areas calculated on the basis both crisp (Eq. 6) and fuzzy distance (Eq. 7) metrics are the same. This fact is straightforwardly deduced from the

definition of influence areas and the crisp and the properties of the fuzzy sets. On the other hand, probabilistic influence areas depend upon initial data crisp or fuzzy partition - the same for the given clusters, but additionally - the other probability parameters - mean values and covariances depend upon the selected data distribution and for this reason, probabilistic influence areas are different for different distribution of data objects.

4.3 RECA-S Set

The RECA-S set in the clustering setting presents the set of the **RECA** cluster centers $C^{reca} = C^r$ together with the points assigned to each of the cluster and RECA parameters RECA-P. RECA Set elements: RECA parameters - **RECA-P**, RECA cluster set - RECA-C. *RECA-S* may be obtained by the procedures of uniform RECA transformations and RECA-S data object assignment.

Definition - RECA N-Blocks - *RECA-S* in this context is understood as a set of objects assigned to the *RECA-C* clusters according to the given parameters set *RECA-S*. Given two different *RECA-S* with the same number of clusters and the same cluster centers *RECA-C*, it is possible to perform the selected set-theoretic operations as described in the next section.

Assignment strategy to RECA clusters. Data points assigned to the cluster C_i^r according to selected assignment strategy S are denoted as $C(a, S)_i^r = \{x_0, x_1, \ldots, x_s\}$. Possible assignment strategies S include arbitrary data objects assignments to RECA components - RECA-C. The RECA component in the clustering setting presents the set of the *RECA-CC* (cluster centers) together with *RECA-CA* (the points assigned to each of the cluster).

Algorithm 1. Different RECA-AS cluster assignment strategies

1. RECA-CC cluster centers obtained during calculation of crisp and fuzzy partitions,
2. RECA-CC cluster centers obtained during calculation of probabilistic and fuzzified probabilistic partitions,
   (a) RECA-P parameters - (C, P, F, FP)(C, P, F, FP)-TD-RECA,
   (b) N cluster centers and covariance matrices, new partitions, new cluster centers,
3. the set of random cluster assignments, new cluster centers RECA-CS,
4. uniform RECA transformations $(C, F, P, FP)(TD)RECA^p$.

The example of *RECA-IA* is presented in Figure 1. In Figure 1 (a) crisp (fuzzy) influence areas have been displayed, in Figure 1 (b) probabilistic influence areas have been displayed.

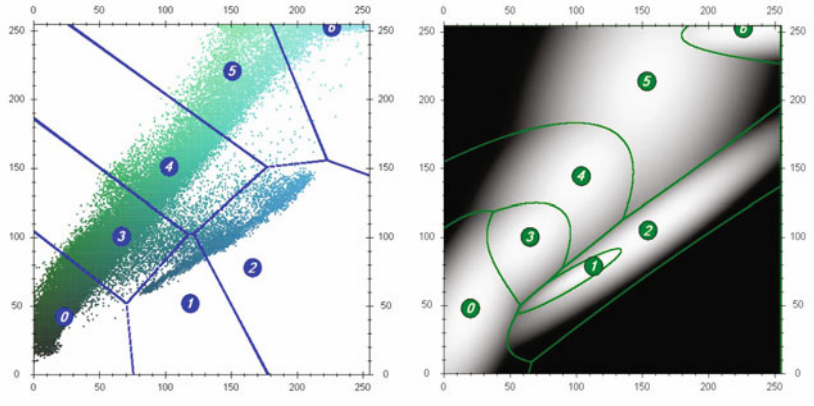


Fig. 1. The color 0GB space with selected (a) crisp (fuzzy) influence areas (b) probabilistic influence areas

5 Set-Theoretic Operations on C-REF Components

5.1 The Procedure of RECA Set Calculation

In standard *RECA* component N cluster centers with points are assigned to each cluster. For each cluster center C_m there are n points assigned to this cluster. *RECA-R* - region determined by the set of *RECA-P* parameters and *RECA-C* clusters in crisp and fuzzy setting. *RECA-R* - region in probabilistic setting is determined by the set of *RECA-P* paramters, *RECA-C* clusters *RECA-S*.

Algorithm 2. RECA-CAS

1. perform the data point assignment, arbitrary or *RECA-S* based point assignment,
2. calculate cluster centers,
3. recalculate point assignment in *RECA-AS* way.

With each cluster, the corresponding RECA-A assigned data object set is defined. RECA-AS consists of the objects that belong to that cluster. Most often RECA-AS is defined by the means of the following criteria.

5.2 RECA Components Operations

The introduction of the notion of the *RECA* component makes possible to perform operations on **RECA** components. Given two different RECA-S with the same number of clusters and the same cluster centers RECA-C, it is possible to perform the following set-theoretic operations *RECA-S* = *RECA-S1*, *RECA-S2*

Algorithm 3. RECA-AS

1. objects in the influence areas of the cluster, crisp or fuzzy IA
2. objects in the influence areas of the RECA-S, RECA-P
3. objects after c(RECA-S)
   - recalculation of cluster centers from RECA-S
   - re-assignment to new c(RECA-S)

1. sum, RECA(S1 ∪ S2)
2. difference, RECA(S1 - S2), RECA(S2 - S1)
3. section, RECA(S1 ∩ S2)
4. symmetric difference, RECA(S1, S2)

The way, the operations are performed in case of the above mentioned definitions, from set-theoretic point of view, it is possible to perform the following steps

1. insert in the result set only data objects with no duplicates,
2. allow to insert duplicate objects in the resultant set, each duplicate object is considered to be unique *sum(S1+S2)* object,
3. allow to insert duplicate objects in the resultant set, each duplicates are considered to be one object, but this object measure is summed up.

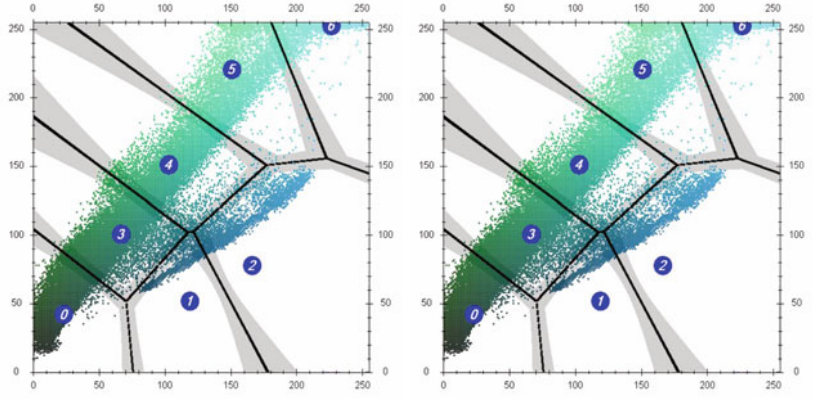


Fig. 2. The color 0GB space with selected influence areas (a) of the cluster set RECA-S with the cluster centers RECA-CC given in (b)

The example of *RECA-IA* is presented in Figure 2. The color *0GB* space with selected influence areas (a) of the cluster set *RECA-S* with the cluster centers *RECA-CC* given in (b).

6 Conclusions and Future Research

In the study, the new algorithmic and analysis clustering framework *C-REF* has been introduced, presented and explained in the setting of image segmentation approach. The introduced *RECA* components create the first starting point for the detailed data analysis routines. The *RECA* components are designed for robust data analysis on the base of crisp, fuzzy, probabilistic and fuzzified probabilistic approaches. The combination of crisp, fuzzy, probabilistic and fuzzified probabilistic rough measures together with application of different distance hypersphere d-metrics seems to be suitable for image segmentation and analysis. Further research directed into extension and incorporation of the introduced *RECA* components in *C-REF* setting in the area of theoretical insight as well as practical applications, for example image analysis routines seems to be a reasonable and fruitful task.

Acknowledgments

The research is supported by the Rector's grant of Biaystok University of Technology.

References

1. Malyszko, D., Stepaniuk, J.: Granular Multilevel Rough Entropy Thresholding in 2D Domain. In: 16th International Conference Intelligent Information Systems, IIS 2008, Zakopane, Poland, June 16-18, pp. 151–160 (2008)
2. Malyszko, D., Stepaniuk, J.: Generalized Hypersphere d-Metric in Rough Measures and Image Analysis. Lectures Notes in Computer Science
3. Małyszko, D., Stepaniuk, J.: Standard and Fuzzy Rough Entropy Clustering Algorithms in Image Segmentation. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 409–418. Springer, Heidelberg (2008)
4. Malyszko, D., Stepaniuk, J.: Adaptive multilevel rough entropy evolutionary thresholding. Information Sciences 180(7), 1138–1158 (2010)
5. Malyszko, D., Stepaniuk, J.: Adaptive Rough Entropy Clustering Algorithms in Image Segmentation. Fundamenta Informaticae 98(2-3), 199–231 (2010)
6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV 2001, vol. (2), pp. 416–423. IEEE Computer Society, Los Alamitos (2001)
7. Pal, S.K., Shankar, B.U., Mitra, P.: Granular computing, rough entropy and object extraction. Pattern Recognition Letters 26(16), 2509–2517 (2005)
8. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
9. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley & Sons, New York (2008)
10. Skowron, A., Stepaniuk, J.: Tolerance Approximation Spaces. Fundamenta Informaticae 27(2-3), 245–253 (1996)
11. Stepaniuk, J.: Rough–Granular Computing in Knowledge Discovery and Data Mining. Springer, Heidelberg (2008)

Granular Representation of Temporal Signals Using Differential Quadratures

Michał Momot[1], Alina Momot[2], Krzysztof Horoba[1], and Janusz Jeżewski[1]

[1] Institute of Medical Technology and Equipment ITAM
Roosevelt 118 Street, 41-800 Zabrze, Poland
[2] Silesian University of Technology, Institute of Computer Science
Akademicka 16 Street, 44-100 Gliwice, Poland
michal.momot@itam.zabrze.pl

Abstract. This article presents the general idea of granular representation of temporal data, particularly signal sampled with constant frequency. The core of presented method is based on using fuzzy numbers as information granules. Three types of fuzzy numbers are considered, as interval numbers, triangular numbers and Gaussian numbers. The input space contains values of first few derivatives of underlying signal, which are computed using certain numerical differentiation algorithms, including polynomial interpolation as well as polynomial approximation. Data granules are constructed using the optimization method according to objective function based on two criteria: high description ability and compactness of fuzzy numbers.

The data granules are subject to the clustering process, namely fuzzy c-means. The centroids of created clusters form a granular vocabulary. Quality of description is quantitatively assessed by reconstruction criterion. Results of numerical experiments are presented, which incorporate exemplary biomedical signal, namely electrocardiographic signal.

1 Introduction

The human-centric approach to data processing [1] includes many cases of granular computing application. Fuzzy sets and fuzzy numbers, in particular, play important role in constructing the data granules for the purpose of temporal signal processing. Certain semantics of descriptors can be built using specific forms of fuzzy numbers, namely interval, triangular or Gaussian ones. The parameters of such fuzzy numbers are determined according to optimization criterion which is based on maximizing objective function. The granulation procedure results in a linguistic description of temporal signals, such as biomedical signal, in a digital form together with quantitative assessment of description quality. The input signals are differentiated in order to capture the changes of signal level, instead of absolute values. However, the differentiating of temporal signal usually leads to amplifying the level of distortion (because of the presence of noise). Various differentiation schemes can be applied there, which allow to estimate derivatives of the underlying function (even up to higher levels) based on a limited knowledge,

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 72–79, 2011.

i.e. values of function in a finite set of arguments (samples). Two contradictory objectives are applied to construct such algorithms. One of them provides high robustness to input errors (noises), which usually accompany a useful part of signal. The second objective is to minimize computational complexity (regarded as time and memory consumption). This enables the numerical differentiation algorithm to operate in a real-time regime.

The aim of this paper is to present method of transforming temporal data, particularly biomedical signal, into data granules set. The presented approach is similar to one previously presented in [4]. However, significant modifications have been made, such as using only changes of signal values (which was achieved by using first few derivatives of underlying signal), using numerical quadratures based on interpolation or approximation polynomial (for the purpose of higher robustness to noise). Also, the assumption has been made, as to avoidance of dependence on any prior information about underlying signal, such as knowledge about temporal or amplitude markers (characteristic points).

The latter part of article is organized as follows. Section 2 contains description of numerical differentiation methods (differential quadratures) in both interpolation and approximation cases. Section 3 presents the main idea of data granules constructing for interval, triangular and Gaussian fuzzy number models. Section 4 presents method of creating granular vocabulary by using fuzzy clustering algorithm together with definition of reconstruction quality measure. Section 5 contains description of numerical experiments and discussion of results obtained. Section 6 contains some implementation remarks. The article ends with a discussion of possible future plans for the development of the presented method.

2 Numerical Differentiation Scheme

Using the signal value changes instead of the values of samples is motivated by a need of making the analysis independent on constant or slowly varying component of signal. This enables the granules to capture the actual nature of analyzed data segments (in non-overlapping time windows). For example, in electrocardiographic signal there is well recognized phenomenon of baseline wander (slow changes of isoelectric line). Several approaches to this problem exist, including base-line shifting, normalization (with constant or adaptively varying scaling factors) [9] or using derivatives of observed signal.

The simple methods of numerical differentiation, such as one-step difference, are usually highly sensitive to the presence of signal disturbances, especially high-frequency noise. This problem could be solved by using more sophisticated algorithms [2,3], including differential quadratures based on polynomial interpolation or approximation [11].

2.1 Differential Quadrature Using Polynomial Interpolation

The proposed signal differentiation method exploits Lagrange polynomial interpolation with equidistant nodes within symmetric fixed-width window. The first

derivative of such polynomial, which can be calculated in an analytical manner, is taken as an estimate of signal derivative. The algorithm of computing this value is briefly described below.

Let R denotes the fixed radius of time window around the Nth sample of signal. The values of interpolated function within this window constitute a vector $\mathbf{s} = [s_{N-R}, \ldots, s_N, \ldots, s_{N+R}]$, the degree of polynomial g is determined by number of samples and it is equal to $2R$, therefore $g(x) = \sum_{j=0}^{2R} a_j\, x^j$ and it is equivalent to vector $[a_0, a_1, \ldots, a_{2R-1}, a_{2R}]$. The nodes of interpolation are equidistant and each node is equivalent to one sample index. This leads to assumption, which can be made without loss of generality, that nodes are of the form:

$$x_{N-R} = -R, \ldots, x_{N-1} = -1, x_N = 0, x_{N+1} = 1, \ldots, x_{N+R} = R \tag{1}$$

and the interpolation method is time-invariant and takes into account only the values of interpolated function in fixed-radius neighborhood. Vector of polynomial coefficients can be easily obtained by solving the system of linear equations $\mathbf{Va} = \mathbf{s}$, where $(\mathbf{V})_{i,j} = i^j$ for $i = -R, \ldots, R$, $j = 0, \ldots, 2R$. Non-singularity of matrix $\mathbf{V}$ leads to solution in the form $\mathbf{a} = \mathbf{V}^{-1}\mathbf{s}$. Moreover, the primary objective of this procedure, being the estimating of function derivative in a center of time window, does not require to determine all polynomial coefficients explicitly, since

$$g'(0) = a_1 = \mathbf{d}^T\mathbf{V}^{-1}\mathbf{s}, \tag{2}$$

where $\mathbf{d} = [0, 1, 0, \ldots, 0]^T$.The differentiation algorithm may be seen as high-pass filter using FIR filter with constant coefficients from the vector $\mathbf{d}^T\mathbf{V}^{-1}$. To calculate the higher order derivatives the above described procedure is repeated iteratively.

2.2 Differential Quadrature Using Polynomial Approximation

Another approach to estimate the value of derivative based on a finite sample set within R radius time window is to choose the polynomial with the degree $D < 2R$. As opposed to the previous case, usually there is no polynomial which takes exact sample values in every nodes. This constitutes an idea of approximating the polynomial coefficients according to some optimality criterion. Using sum of squared differences leads to following objective function:

$$G(\mathbf{a}) = \sum_{j=-R}^{R} \left(s_j - \sum_{i=0}^{D} a_i j^i, \right)^2 \tag{3}$$

which after differentiating leads to following formula:

$$\frac{\partial\, G_k(\mathbf{a})}{\partial\, a_n} = -2 \sum_{j=-R}^{R} \left(s_j - \sum_{i=0}^{D} a_i j^i \right) j^n. \tag{4}$$

In order to minimize the objective function the following system of linear equations need to be solved:

$$\sum_{i=0}^{D} a_i \left(\sum_{j=-R}^{R} j^{n+i} \right) = \sum_{j=-R}^{R} j^n s_j \quad \text{for} \quad n = 0, \ldots, D. \tag{5}$$

Similarly as in the case of polynomial interpolation, the estimate of function derivative may be obtained from formula

$$g'(0) = \mathbf{d}^{\mathrm{T}} \mathbf{J}^{-1} \mathbf{y} = \mathbf{d}^{\mathrm{T}} \mathbf{J}^{-1} \mathbf{H} \mathbf{s}, \tag{6}$$

where $(\mathbf{J})_{i,j} = \sum_{k=-R}^{R} k^{i+j}$ for $i, j = 0, \ldots, D$, $(\mathbf{H})_{i,j} = j^i$ for $i = 0, \ldots, D$, $j = -R, \ldots, R$ and $\mathbf{d} = [0, 1, 0, \ldots, 0]^T$.

3 Determining Parameters of Data Granules

The general idea of proposed granulation methods consists in splitting temporal signal s into non-overlapping segments $[s_{iL}, s_{iL+1}, \ldots, s_{(i+1)L-1}]$ with fixed length L. For ith segment the data granule is constructed as a fuzzy number based on the values of signal amplitude changes (expressed in its derivatives). The choice of symmetric fuzzy number models (interval, triangular and Gaussian) was motivated by their unimodal membership functions:

$$\mu_{c,r}(t) = \begin{cases} 1 & \text{if } t \in (c-r, c+r) \\ 0 & \text{if } t \notin (c-r, c+r) \end{cases} \qquad \text{(interval memb. func.)}, \tag{7}$$

$$\mu_{c,r}(t) = \left\{ 1 - \frac{|t-c|}{r}, 0 \right\} \qquad \text{(triang. memb. func.)}, \tag{8}$$

$$\mu_{c,r}(t) = \exp\left\{ -\frac{(t-c)^2}{r} \right\} \qquad \text{(Gauss. memb. func.)}, \tag{9}$$

where c is the center of fuzzy number and r is radius parameter, determining the support of its membership function (in case of interval and triangular models) or the dispersion parameter (in case of Gaussian model the support is unbounded). The center of fuzzy number is equal to median of values in corresponding data segment. The choice of median as a location parameter was justified by its robustness to outliers in the input data. Assuming the length of data window to be an odd number

$$c_i^{(k)} = \tilde{s}_{iL+\frac{L-1}{2}}^{(k)}, \tag{10}$$

where $c_i^{(k)}$ denotes center of fuzzy number, $[\tilde{s}_{iL}, \tilde{s}_{iL+1}, \ldots, \tilde{s}_{(i+1)L-1}]$ is ith segment of kth order data derivatives sorted in ascending order. The choice of radius (or dispersion) parameters is determined by two contradictory requirements: maximizing the empirical evidence of the input data while making the granule

most specific. The first criterion leads to maximizing the overall membership for data segment:

$$\sum_{\substack{j=0 \\ j \neq iL+\frac{L-1}{2}}}^{L-1} \mu_{c_i^k,r}(\tilde{s}_{iL+j}^{(k)}), \tag{11}$$

excluding the membership value for the central sample (median), because it does not add any significant information. The second criterion calls for maximum specificity of fuzzy number, which is expressed by minimizing radius (or dispersion) parameter r. These two opposing criteria lead to following objective function:

$$G(r) = r^{-1} \sum_{\substack{j=0 \\ j \neq iL+\frac{L-1}{2}}}^{L-1} \mu_{c_i^k,r}(\tilde{s}_{iL+j}^{(k)}). \tag{12}$$

Parameters of fuzzy numbers, being centers and radii (or dispersion parameters), are determined for every segment and for signal derivatives up to previously fixed order K, which leads to following granular description: $[c_i^{(1)}, r_i^{(1)}, \ldots, c_i^{(K)}, r_i^{(K)}]$.

4 Granular Vocabulary Construction

The granular representation of data segments, being collection of $2K$-dimensional vectors is subject to clustering process, which leads to determining the descriptors that reflects the structure of whole set of granules. In proposed approach the well known fuzzy c-means (FCM) method is used [2][7]. The main idea of FCM is to determine, given the collection of N input data in form of D-dimensional vectors, the membership values that minimize the following objective function

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{i,k}^m \|\mathbf{z}_k - \mathbf{v}_i\|^2, \tag{13}$$

where $\mathbf{v}_i$ are cluster prototypes (centroids), $\|\cdot\|$ denotes some norm in D-dimensional space (usually Euclidean norm) and $m > 1$ is fuzziness parameter controlling impact of the membership grades on individual clusters. The clustering procedure is performed in an iterative manner, alternating between computing the membership values and updating cluster prototypes, until some termination condition is fulfilled. After determining the cluster prototypes coordinates, the following formula enables to reconstruct input data vector based on obtained granular vocabulary:

$$h(\mathbf{z}) = \frac{\sum_{i=1}^{c} \left(u_i(\mathbf{z})\right)^m \mathbf{v}_i}{\sum_{i=1}^{c} \left(u_i(\mathbf{z})\right)^m}, \tag{14}$$

where $u_i(\mathbf{z})$ is the degree of membership in ith cluster for the given input vector $\mathbf{z}$. This leads to the reconstruction criterion

$$V = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{z}_k - h(\mathbf{z}_k)\|^2 \tag{15}$$

5 Numerical Experiments

The numerical experiments were performed in order to evaluate the performance of proposed method empirically. The input data were a part of MIT-BIH database [8][9]. This set of data contains excerpts from two-channel ambulatory electrocardiographic recordings. For the purpose of this study over 10 minutes signal has been considered (the recordings came from different patients). Only single channel signal was taken into account and the tests did not incorporate any prior information, like human- or computer-readable annotations. Derivatives of first and second order were calculated ($K = 2$) with the approximation window radius $R = 4$ and degree of polynomial $D = 3$ (on the basis of visual assessment of the quality of the signal derivative). The granulation window length L varied from 5 to 11 (only odd numbers) and the maximum number of clusters was set to 6. The input values of signal, as well as results of computation were expressed in μV. The procedure was repeated for all considered types of fuzzy numbers. In all cases the clustering was performed and the reconstruction ability was empirically evaluated. For every pair of L and c the empirical reconstruction error was presented as a function of fuzziness parameter m (in range 1-3). The optimal values m_{opt} of parameter m was determined, which resulted in minimal reconstruction error V_{opt}. In general, increasing c and L resulted in better reconstruction ability. There is also the interesting phenomenon: in most cases the reconstruction error increases rapidly for m exceeding some threshold value m_{thresh}. However, for $L > 7$ and $c > 3$ the functions become more smooth and the point of rapid change moves to the right ($m_{thresh} > 2$). The general shape of reconstruction error as a function of m was the same for all types of fuzzy numbers (interval, triangular and Gaussian). It suggests that the method will work stable for selected values of parameter m, namely in a range $1.0 < m < 2.0$. Table 1 presents exact empirical values of m_{opt}, V_{opt} and m_{thresh} for selected values of c and L in case of triangular fuzzy numbers.

6 Implementation Remarks

It is worth making a mention about how the described method has been implemented. Since the presented algorithms have high computational complexity, there is a need for techniques to reduce computation time. It can be easily observed that processes of constructing data granules in separate granulation windows are performed independently from each other, therefore they can be performed concurrently. The use of technology CUDA (Compute Unified Device Architecture, [5]) at this stage made it possible to significantly reduce computation time.

Table 1. Optimal empirical values of reconstruction criterion and corresponding parameters in case of triangular fuzzy number

L	c	V_{opt}	m_{opt}	m_{thresh}
5	2	1569.15	1.150	1.225
	3	533.28	1.650	1.650
	4	484.26	1.050	1.725
	5	374.97	1.350	1.975
	6	350.20	1.625	2.050
7	2	944.37	1.175	1.325
	3	311.03	1.775	1.775
	4	285.93	1.725	2.000
	5	235.96	1.500	2.100
	6	209.52	1.325	2.200
9	2	432.81	1.300	1.900
	3	201.48	1.900	1.925
	4	167.19	1.225	2.575
	5	148.65	1.150	2.750
	6	116.89	1.450	2.875
11	2	166.63	1.575	-
	3	102.27	1.700	1.525
	4	87.16	1.625	2.175
	5	70.89	1.550	2.350
	6	58.13	1.550	2.425

7 Conclusion

The issues described in this article will be the subject of further investigation. It is planned to develop a modified and extended version of the presented method, especially considering modified optimality criteria for the construction of data granules, as well as other clustering algorithms. Some other plans relate to applications of presented method. Since, as has been mentioned previously, the granulation method itself does not assume knowledge about any characteristic points of analyzed signal, it is possible to apply it to determine the linguistic description function of temporal signals.

Acknowledgements

This research was partially supported by Polish Ministry of Science and Higher Education as Research Project N N518 406438.

References

1. Bargiela, A., Pedrycz, W.: Granular Computing: An Introduction. Kluwer Academic Publishers, Dordercht (2003)
2. Braci, M., Diop, S.: On numerical differentiation algorithms for nonlinear estimation. In: Proc. 42nd IEEE Conf. on Decision and Control, Maui, Hawaii, USA, pp. 2896–2901 (2003)

3. Burden, R.L., Faires, J.D.: Numerical Analysis. Brooks/Cole, Monterey (2000)
4. Gacek, A., Pedrycz, W.: A Granular Description of ECG Signals. IEEE Trans. Biomed. Eng. 53(10), 1972–1982 (2006)
5. Garland, M., et al.: Parallel Computing Experiences with CUDA. IEEE Micro 28(4), 13–27 (2008)
6. Hathaway, R., Bezdek, J., Hu, Y.: Generalized fuzzy c-means clustering strategies using Lp norm distances. IEEE Trans. Fuzzy Syst. 8(5), 576–582 (2000)
7. Mark, R., Moody, G.: MIT-BIH Arrhythmia Database Directory. MIT, Cambridge (1988)
8. Moody, G.B., Mark, R.G.: The MIT-BIH arrhythmia database on CD-ROM and software for use with it. In: Proc. Conf. Computers in Cardiology, San Diego, CA, pp. 185–188 (1990)
9. Ortolani, M., Hofer, H., et al.: Fuzzy Information Granules in Time Series Data. In: Proc. IEEE Int. Conf. on Fuzzy Systems, pp. 695–699 (2002)
10. Pedrycz, W.: Fuzzy Sets as a User-Centric Processing Framework of Granular Computing. In: Pedrycz, W., Skowron, A., Kreinovich, V. (eds.) Handbook of Granular Computing, pp. 97–139. John Wiley & Sons, Chichester (2008)
11. Zhang, Y.: Advanced Differential Quadrature Methods. CRC Press, Boca Raton (2009)

An Adjustable Approach to Interval-Valued Intuitionistic Fuzzy Soft Sets Based Decision Making

Hongwu Qin, Xiuqin Ma, Tutut Herawan, and Jasni Mohamad Zain

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
qhwump@gmail.com, xueener@yahoo.com.cn,
tutut@ump.edu.my, jasni@ump.edu.my

Abstract. Research on soft set based decision making has received much attention in recent years. Feng et al. presented an adjustable approach to fuzzy soft sets based decision making by using level soft set, and subsequently extended the approach to interval-valued fuzzy soft set based decision making. Jiang et al. generalize the approach to solve intuitionistic fuzzy soft sets based decision making. In this paper, we further generalize the approaches introduced by Feng et al. and Jiang et al. Using reduct intuitionistic fuzzy soft sets and level soft sets of intuitionistic fuzzy soft sets, an adjustable approach to interval-valued intuitionistic fuzzy soft set based decision making is presented. Some illustrative example is employed to show the feasibility of our approach in practical applications.

Keywords: Soft sets, Interval-valued intuitionistic fuzzy soft sets, Decision making, Level soft sets.

1 Introduction

In 1999, Molodtsov [1] proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. At present, work on the soft set theory is progressing rapidly and many important theoretic results have been achieved [2-8].

The research on fuzzy soft set has also received much attention since its introduction by Maji et al. [9]. Majumdar and Samanta [10] have further generalized the concept of fuzzy soft sets. Maji et al. [11-13] extended soft sets to intuitionistic fuzzy soft sets. Yang et al. [14] proposed the concept of the interval-valued fuzzy soft sets. Jiang et al. [15] proposed a more general soft set model called interval-valued intuitionistic fuzzy soft set by combining the interval-valued intuitionistic fuzzy sets and soft sets.

At the same time, there has been some progress concerning practical applications of soft set theory, especially the use of soft sets in decision making [16-21]. Maji et al. [16] first applied soft sets to solve the decision making problem. Roy and Maji [17] presented a novel method to cope with fuzzy soft sets based decision making problems. Kong et al. [18] pointed out that the Roy-Maji method [17] was incorrect and they presented a revised algorithm. Feng et al. [19] discussed the validity of the

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 80–89, 2011.

Roy-Maji method [17] and presented an adjustable approach to fuzzy soft sets based decision making by means of level soft sets. Yang et al. [14] applied the interval-valued fuzzy soft sets to analyze a decision making problem. The method they used is based on fuzzy choice value. Feng et al. [20] gave deeper insights into decision making involving interval-valued fuzzy soft sets. They analyzed the inherent drawbacks of fuzzy choice value based method and proposed a flexible scheme by using reduct fuzzy soft sets and level soft sets. Similarly, Jiang et al. [21] presented an adjustable approach to intuitionistic fuzzy soft sets based decision making by using level soft sets of intuitionistic fuzzy soft sets.

In this paper, we further generalize the approaches introduced by Feng et al. [20] and Jiang et al. [21]. Concretely, we define the notion of reduct intuitionistic fuzzy soft sets and present an adjustable approach to interval-valued intuitionistic fuzzy soft set based decision making by using reduct intuitionistic fuzzy soft sets and level soft sets of intuitionistic fuzzy soft sets. Firstly, by computing the reduct intuitionistic fuzzy soft set, an interval-valued intuitionistic fuzzy soft set is converted into an intuitionistic fuzzy soft set, and then the intuitionistic fuzzy soft set is further converted into a crisp soft set by using level soft sets of intuitionistic fuzzy soft sets. Finally, decision making is performed on the crisp soft set.

The rest of this paper is organized as follows. The following section briefly reviews some basic notions of soft sets. Section 3.1 defines the concept of reduct intuitionistic fuzzy soft sets. Section 3.2 recalls the level soft sets. We present our algorithm to interval-valued intuitionistic fuzzy soft set based decision making problems and illustrate example in Section 3.3. Finally, conclusions are given in Section 4.

2 Preliminaries

In this section, we recall the basic notions of soft sets, interval-valued fuzzy soft sets and interval-valued intuitionistic fuzzy soft sets.

Let U be an initial universe of objects, E be the set of parameters in relation to objects in U, $\mathscr{P}(U)$ denote the power set of U and $A \subseteq E$. The definition of soft set is given as follows.

Definition 2.1 ([1]). A pair (F, A) is called *a soft set* over U, where F is a mapping given by

$$F : A \to \mathscr{P}(U).$$

From definition, a soft set (F, A) over the universe U is a parameterized family of subsets of the universe U, which gives an approximate description of the objects in U. Before introduce the notion of the interval-valued fuzzy soft sets, let us give the concept of the interval-valued fuzzy sets [22].

An interval-valued fuzzy set Y on an universe U is a mapping such that $Y : U \to Int([0,1])$, where $Int([0,1])$ stands for the set of all closed subintervals of [0, 1], the set of all interval-valued fuzzy sets on U is denoted by $\mathscr{I}(U)$. Suppose that

$Y \in \mathscr{J}(U), \forall x \in U$, $Y(x) = [Y^-(x), Y^+(x)]$ is called the degree of membership an element x to Y. $Y^-(x)$ and $Y^+(x)$ are referred to as the lower and upper degrees of membership x to Y, where $0 \le Y^-(x) \le Y^+(x) \le 1$.

Definition 2.2 ([14]). Let U be an initial universe and E be a set of parameters, $A \subseteq E$, a pair (F, A) is called *an interval-valued fuzzy soft set* over U, where F is a mapping given by

$$F : A \to \mathscr{J}(U).$$

An interval-valued fuzzy soft set is a parameterized family of interval-valued fuzzy subsets of U. $\forall \varepsilon \in A$, $F(\varepsilon)$ is referred as the interval fuzzy value set of paramete ε. Clearly, $F(\varepsilon)$ can be written as $F(\varepsilon) = \{\langle x, F(\varepsilon)^-(x), F(\varepsilon)^+(x)\rangle : x \in U\}$, where $F(\varepsilon)^-(x)$ and $F(\varepsilon)^+(x)$ be the lower and upper degrees of membership of x to $F(\varepsilon)$ respectively.

Finally, we will introduce the concepts of interval-valued intuitionistic fuzzy set [23] and interval-valued intuitionistic fuzzy soft set.

An interval-valued intuitionistic fuzzy set on a universe Y is an object of the form $A = \{\langle x, \mu_A(x), \gamma_A(x)\rangle \mid x \in Y\}$, $\mu_A(x) : Y \to Int([0,1])$ and $\gamma_A(x) : Y \to Int([0,1])$ satisfy the following condition: $\forall x \in Y$, $\sup \mu_A(x) + \sup \gamma_A(x) \le 1$.

Definition 2.3 ([15]). Let U be a universe set, E be a set of parameters, $\mathscr{E}(U)$ denotes the set of all interval-valued intuitionistic fuzzy sets of U and $A \subseteq E$. A pair (F, A) is called *an interval-valued intuitionistic fuzzy soft set* over U, where F is a mapping given by

$$F : A \to \mathscr{E}(U).$$

$\forall \varepsilon \in A$, $F(\varepsilon)$ is an interval-valued intuitionistic fuzzy set of U. $F(\varepsilon)$ can be written as: $F(\varepsilon) = \{\langle x, \mu_{F(\varepsilon)}(x), \gamma_{F(\varepsilon)}(x)\rangle : x \in U\}$, where $\mu_{F(\varepsilon)}(x)$ is the interval-valued fuzzy membership degree, $\gamma_{F(\varepsilon)}(x)$ is the interval-valued fuzzy non-membership degree. For illustration, we consider the following house example.

Example 1. Consider an interval-valued intuitionistic fuzzy soft set (F, A) which describes the "attractiveness of houses" that Mr. X is considering for purchase. Suppose that there are six houses under consideration, namely the universes $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$, and the parameter set $A = \{e_1, e_2, e_3, e_4, e_5\}$, where e_i stand for "beautiful", "large", "cheap", "modern" and "in green surroundings" respectively. The tabular representation of (F, A) is shown in Table 1. Obviously, we can see that the precise evaluation for each object on each parameter is unknown while the lower and upper limits of such an evaluation are given. For example, we cannot present the precise membership degree and non-membership degree of how beautiful house h_1 is,

however, house h_1 is at least beautiful on the membership degree of 0.7 and it is at most beautiful on the membership degree of 0.8; house h_1 is not at least beautiful on the non-membership degree of 0.1 and it is not at most beautiful on the non-membership degree of 0.2.

Table 1. Interval-valued intuitionistic fuzzy soft set (F, A)

U	e_1	e_2	e_3	e_4	e_5
h_1	[0.7,0.8], [0.1,0.2]	[0.82,0.84], [0.05,0.15]	[0.52,0.72], [0.18,0.25]	[0.55,0.6], [0.3,0.35]	[0.7,0.8], [0.1,0.2]
h_2	[0.85,0.9], [0.05,0.1]	[0.7,0.74], [0.17,0.25]	[0.7,0.75], [0.1,0.23]	[0.7,0.75], [0.15,0.25]	[0.75,0.9], [0.05,0.1]
h_3	[0.5,0.7], [0.2,0.3]	[0.86,0.9], [0.04,0.1]	[0.6,0.7], [0.2,0.28]	[0.2,0.3], [0.5,0.6]	[0.65,0.8], [0.15,0.2]
h_4	[0.4,0.6], [0.3,0.4]	[0.52,0.64], [0.23,0.35]	[0.72,0.78], [0.11,0.21]	[0.3,0.5], [0.4,0.5]	[0.8,0.9], [0.05,0.1]
h_5	[0.6,0.8], [0.15,0.2]	[0.3,0.35], [0.5,0.65]	[0.58,0.68], [0.18,0.3]	[0.68,0.77], [0.1,0.2]	[0.72,0.85], [0.1,0.15]
h_6	[0.3,0.5], [0.3,0.45]	[0.5,0.68], [0.25,0.3]	[0.33,0.43], [0.5,0.55]	[0.62,0.65], [0.15,0.35]	[0.84,0.93], [0.04,0.07]

3 An Adjustable Approach to Interval-Valued Intuitionistic Fuzzy Soft Set Based Decision Making

In this section we present an adjustable approach to interval-valued intuitionistic fuzzy soft set based decision making problems by combining the reduct intuitionistic fuzzy soft sets and level soft sets of intuitionistic fuzzy soft sets. First we define the concept of reduct intuitionistic fuzzy soft sets, and then recall the level soft sets, finally present our algorithm and illustrate example.

3.1 Reduct Intuitionistic Fuzzy Soft Sets

The concept of reduct fuzzy soft set is proposed in [20]. By adjusting the value of opinion weighting vector, an interval-valued fuzzy soft set can be converted into a fuzzy soft set, which makes the making decision based on interval-valued fuzzy soft set much easier.

Similarly, we can introduce the idea to making decision based on interval-valued intuitionistic fuzzy soft set, that is, convert both interval-valued membership degree and interval-valued non-membership degree into one fuzzy value. As a result, an interval-valued intuitionistic fuzzy soft set will be transformed to an intuitionistic fuzzy soft set, which will facilitate the making decision based on interval-valued intuitionistic fuzzy soft set. We define the notion of reduct intuitionistic fuzzy soft set as follows to illustrate the idea.

Let U be a universe set, E be a set of parameters and $A \subseteq E$. Let (F, A) be an interval-valued intuitionistic fuzzy soft set over U such that $\forall \varepsilon \in A$, $F(\varepsilon)$ is an interval-valued intuitionistic fuzzy set with $F(\varepsilon) = \{\langle x, \mu_{F(\varepsilon)}(x), \gamma_{F(\varepsilon)}(x)\rangle : x \in U\}$, $\forall x \in U$.

Definition 3.1. Let $\alpha, \beta, \phi, \varphi \in [0,1]$, $\alpha + \beta = 1$, $\phi + \varphi = 1$. The vector $W = (\alpha, \beta, \phi, \varphi)$ is called an opinion weighting vector. The intuitionistic fuzzy soft set (F_W, A) over U such that

$$F_W(\varepsilon) = \{(x, \alpha\mu_{F(\varepsilon)}{}^{-}(x) + \beta\mu_{F(\varepsilon)}{}^{+}(x), \phi\gamma_{F(\varepsilon)}{}^{-}(x) + \varphi\gamma_{F(\varepsilon)}{}^{+}(x)) : x \in U\}, \forall \varepsilon \in A,$$

is called the *weighted reduct intuitionistic fuzzy soft set* of the interval-valued intuitionistic fuzzy soft set (F, A) with respect to the opinion weighting vector W.

By adjusting the value of α, β, ϕ and φ, an interval-valued intuitionistic fuzzy soft set can be converted into any reduct intuitionistic fuzzy soft set decision maker desired. Specially, let $\alpha = 1$, $\beta = 0$, $\phi = 0$ and $\varphi = 1$, we have the pessimistic-pessimistic reduct intuitionistic fuzzy soft set (PPRIFS), denoted by (F_{-+}, A) and defined by

$$F_{-+}(\varepsilon) = \{(x, \mu_{F(\varepsilon)}{}^{-}(x), \gamma_{F(\varepsilon)}{}^{+}(x)) : x \in U\}, \forall \varepsilon \in A.$$

Let $\alpha = 0$, $\beta = 1$, $\phi = 1$ and $\varphi = 0$, we have the optimistic-optimistic reduct intuitionistic fuzzy soft set (OORIFS), denoted by (F_{+-}, A) and defined by

$$F_{+-}(\varepsilon) = \{(x, \mu_{F(\varepsilon)}{}^{+}(x), \gamma_{F(\varepsilon)}{}^{-}(x)) : x \in U\}, \forall \varepsilon \in A.$$

Let $\alpha = 0.5$, $\beta = 0.5$, $\phi = 0.5$ and $\varphi = 0.5$, we have the neutral-neutral reduct intuitionistic fuzzy soft set (NNRIFS), denoted by (F_{NN}, A) and defined by

$$F_{NN}(\varepsilon) = \{(x, (\mu_{F(\varepsilon)}{}^{-}(x) + \mu_{F(\varepsilon)}{}^{+}(x))/2, (\gamma_{F(\varepsilon)}{}^{-}(x) + \gamma_{F(\varepsilon)}{}^{+}(x))/2) : x \in U\}, \forall \varepsilon \in A.$$

Example 2. Compute the PPRIFS (F_{-+}, A), OORIFS (F_{+-}, A) and NNRIFS (F_{NN}, A) of the interval- valued intuitionistic fuzzy soft set (F, A) shown in Table 1. The results are shown in Table 2, 3 and 4 respectively.

Table 2. Pessimistic-pessimistic reduct intuitionistic fuzzy soft set of (F, A)

U	e_1	e_2	e_3	e_4	e_5
h_1	[0.7, 0.2]	[0.82, 0.15]	[0.52, 0.25]	[0.55, 0.35]	[0.7,0.2]
h_2	[0.85, 0.1]	[0.7, 0.25]	[0.7, 0.23]	[0.7, 0.25]	[0.75, 0.1]
h_3	[0.5, 0.3]	[0.86, 0.1]	[0.6, 0.28]	[0.2, 0.6]	[0.65, 0.2]
h_4	[0.4, 0.4]	[0.52,0.35]	[0.72,0.21]	[0.3, 0.5]	[0.8, 0.1]
h_5	[0.6, 0.2]	[0.3, 0.65]	[0.58, 0.3]	[0.68, 0.2]	[0.72, 0.15]
h_6	[0.3, 0.45]	[0.5, 0.3]	[0.33, 0.55]	[0.62, 0.35]	[0.84, 0.07]

Table 3. Optimistic-optimistic reduct intuitionistic fuzzy soft set of (F, A)

U	e_1	e_2	e_3	e_4	e_5
h_1	[0.8, 0.1]	[0.84, 0.05]	[0.72, 0.18]	[0.6, 0.3]	[0.8,0.1]
h_2	[0.9, 0.05]	[0.74, 0.17]	[0.75, 0.1]	[0.75, 0.15]	[0.9, 0.05]
h_3	[0.7, 0.2]	[0.9, 0.04]	[0.7, 0.2]	[0.3, 0.5]	[0.8, 0.15]
h_4	[0.6, 0.3]	[0.64,0.23]	[0.78,0.11]	[0.5, 0.4]	[0.9, 0.05]
h_5	[0.8, 0.15]	[0.35, 0.5]	[0.68, 0.18]	[0.77, 0.1]	[0.85, 0.1]
h_6	[0.5, 0.3]	[0.68, 0.25]	[0.43, 0.5]	[0.65, 0.15]	[0.93, 0.04]

Table 4. Neutral-neutral reduct intuitionistic fuzzy soft set of (F, A)

U	e_1	e_2	e_3	e_4	e_5
h_1	[0.75, 0.15]	[0.83, 0.1]	[0.62, 0.22]	[0.58, 0.33]	[0.75,0.15]
h_2	[0.88, 0.08]	[0.72, 0.21]	[0.73, 0.17]	[0.73, 0.2]	[0.83, 0.08]
h_3	[0.6, 0.25]	[0.88, 0.07]	[0.65, 0.24]	[0.25, 0.55]	[0.73, 0.18]
h_4	[0.5, 0.35]	[0.58,0.29]	[0.75,0.16]	[0.4, 0.45]	[0.85, 0.08]
h_5	[0.7, 0.18]	[0.33, 0.58]	[0.63, 0.24]	[0.73, 0.15]	[0.79, 0.13]
h_6	[0.4, 0.38]	[0.59, 0.28]	[0.38, 0.53]	[0.64, 0.25]	[0.89, 0.06]

3.2 Level Soft Sets

Feng et al. [19] initiated the concept of level soft set to solve fuzzy soft set based decision making problem. Subsequently, the same author applied level soft set to solve interval-valued fuzzy soft sets based decision making problem [20]. Jiang et al. [21] further generalize the approach introduced in [19] by applying level soft set to solve intuitionistic fuzzy soft sets based decision making. Level soft set of intuitionistic fuzzy soft set is defined as follows.

Definition 3.2 ([21]). Let $\varpi = (F, A)$ be an intuitionistic fuzzy soft set over U, where $A \subseteq E$ and E is a set of parameters. Let $\lambda : A \to [0,1]\times[0,1]$ be an intuitionistic fuzzy set in A which is called a threshold intuitionistic fuzzy set. The level soft set of ϖ with respect to λ is a crisp soft set $L(\varpi;\lambda) = (F_\lambda, A)$ defined by

$$F_\lambda(\varepsilon) = L(F(\varepsilon);\lambda(\varepsilon)) = \{x \in U \mid \mu_{F(\varepsilon)}(x) \geq \mu_\lambda(\varepsilon) \quad \text{and} \quad \gamma_{F(\varepsilon)}(x) \leq \gamma_\lambda(\varepsilon)\} \quad \forall \varepsilon \in A.$$

According to the definition, four types of special level soft set are also defined in [21], which are called Mid-level soft set $L(\varpi, mid_\varpi)$, Top-Bottom-level soft set $L(\varpi, topbottom_\varpi)$, Top-Top-level soft set $L(\varpi, toptop_\varpi)$ and Bottom-bottom-level soft set $L(\varpi, bottombottom_\varpi)$.

3.3 Our Algorithm for Decision Making Based on Interval-Valued Intuitionistic Fuzzy Soft Sets

In this section we present our algorithm for decision making based on interval-valued intuitionistic fuzzy soft sets. By considering appropriate reduct intuitionistic fuzzy

soft sets and level soft sets of intuitionistic fuzzy soft sets, interval-valued intuitionistic fuzzy soft sets based decision making can be converted into only crisp soft sets based decision making. Firstly, by computing the reduct intuitionistic fuzzy soft set, an interval-valued intuitionistic fuzzy soft set is converted into an intuitionistic fuzzy soft set, and then the intuitionistic fuzzy soft set is further converted into a crisp soft set by using level soft sets of intuitionistic fuzzy soft sets. Finally, decision making is performed on the crisp soft set. The details of our algorithm are listed below.

Algorithm 1.

1. Input the interval-valued intuitionistic fuzzy soft set (F, A).
2. Input an opinion weighting vector $W = (\alpha, \beta, \phi, \varphi)$ and compute the weighted reduct intuitionistic fuzzy soft set $\varpi = (F_W, A)$ of the interval-valued intuitionistic fuzzy soft set (F, A) with respect to the opinion weighting vector W (or choose $\varpi =$ PPRIFS (F_{-+}, A), OORIFS (F_{+-}, A) or NNRIFS (F_{NN}, A) of (F, A)).
3. Input a threshold intuitionistic fuzzy set $\lambda : A \to [0,1]\times[0,1]$ (or give a threshold value pair $(s,t) \in [0,1]\times[0,1]$); or choose the mid-level decision rule; or choose the top-bottom-level decision rule; or choose the top-top-level decision rule; or choose the bottom-bottom-level decision rule) for decision making.
4. Compute the level soft set $L(\varpi; \lambda)$ with regard to the threshold intuitionistic fuzzy set λ (or the (s,t)-level soft set $L(\varpi; s,t)$; or the mid-level soft set $L(\varpi, mid_{\varpi})$;or the top-bottom-level soft set $L(\varpi, topbottom_{\varpi})$;or the top-top-level soft set $L(\varpi, toptop_{\varpi})$; or the bottom-bottom-level soft set $L(\varpi, bottombottom_{\varpi})$).
5. Present the level soft set $L(\varpi; \lambda)$ (or $L(\varpi; s,t)$; or $L(\varpi, mid_{\varpi})$; or $L(\varpi, topbottom_{\varpi})$; or $L(\varpi, toptop_{\varpi})$; or $L(\varpi, bottombottom_{\varpi})$)in tabular form and compute the choice value c_i of o_i, $\forall i$.
6. The optimal decision is to select o_k if $c_k = \max_i c_i$.
7. If k has more than one value then any one of o_k may be chosen.

There are two remarks here.

Firstly, reader is referred to [16] for more details regarding the method of computing the choice value in the fifth step of the above algorithm,

Secondly, in the last step of Algorithm 1, one may go back to the step 2 or step 3 to modify opinion weighting vector or the threshold so as to adjust the final optimal decision when there are too many "optimal choices" to be chosen.

The advantages of Algorithm 1 are mainly twofold.

Firstly, we need not treat interval-valued intuitionistic fuzzy soft sets directly in decision making but only deal with the related reduct intuitionistic soft sets and finally the crisp level soft sets after choosing certain opinion weighting vectors and thresholds. This makes our algorithm simpler and easier for application in practical problems.

Secondly, there are a large variety of opinion weighting vectors and thresholds that can be used to find the optimal choices, hence our algorithm has great flexibility and

adjustable capability. Table 5 gives some typical schemes that arise from Algorithm 1 by combining reduct intuitionistic soft set PPRIFS (F_{-+}, A) and several typical level soft sets. As pointed out in [19], many decision making problems are essentially humanistic and subjective in nature; hence there actually does not exist a unique or uniform criterion for decision making in an imprecise environment. This adjustable feature makes Algorithm 1 not only efficient but more appropriate for many practical applications.

To illustrate the basic idea of Algorithm 1, let us consider the following example.

Example 3. Let us reconsider the decision making problem based on the interval-valued intuitionistic fuzzy soft sets (F, A) as in Table 1.

If we select the first scheme "Pes-Mid" in Table 5 to solve the problem, at first we compute the reduct intuitionistic fuzzy soft set PPRIFS $\varpi = (F_{-+}, A)$ as in Table 2 and then use the mid-level decision rule on $\varpi = (F_{-+}, A)$ and obtain the mid-level soft set $L(\varpi; mid)$ with choice values as in Table 6.

From Table 6, it is clear that the maximum choice value is $c_2 = 5$, so the optimal decision is to select h_2.

Table 5. Typical schemes for interval-valued intuitionistic fuzzy soft set based decision making

Scheme	Reduct intuitionistic fuzzy soft set	Level soft set
Pes-Topbot	PPRIFS $\varpi = (F_{-+}, A)$	$L(\varpi; topbottom)$
Pes-Toptop	PPRIFS $\varpi = (F_{-+}, A)$	$L(\varpi; toptop)$
Pes-Mid	PPRIFS $\varpi = (F_{-+}, A)$	$L(\varpi; mid)$
Pes-Botbot	PPRIFS $\varpi = (F_{-+}, A)$	$L(\varpi; bottombottom)$

Table 6. Tabular representation of the mid-level soft set $L(\varpi; mid)$ with choice values

U	e_1	e_2	e_3	e_4	e_5	Choice value(c_i)
h_1	1	1	0	1	0	$c_1 = 3$
h_2	1	1	1	1	1	$c_2 = 5$
h_3	0	1	1	0	0	$c_3 = 2$
h_4	0	0	1	0	1	$c_4 = 2$
h_5	1	0	1	1	0	$c_5 = 3$
h_6	0	0	0	1	1	$c_6 = 2$

4 Conclusion

In this paper, we present an adjustable approach to interval-valued intuitionistic fuzzy soft set based decision making by using reduct intuitionistic fuzzy soft sets and level

soft sets of intuitionistic fuzzy soft sets, which further generalize the approaches introduced by Feng et al. [20] and Jiang et al. [21]. An interval-valued intuitionistic fuzzy soft set based decision making problem is converted into a crisp soft set based decision making problem after choosing certain opinion weighting vectors and thresholds. This makes our algorithm simpler and easier for application in practical problems. In addition, a large variety of opinion weighting vectors and thresholds that can be used to find the optimal alternatives make our algorithm more flexible and adjustable.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Molodtsov, D.: Soft set theory_First results. Computers and Mathematics with Applications 37, 19–31 (1999)
2. Maji, P.K., Biswas, R., Roy, A.R.: Soft set theory. Computers and Mathematics with Applications 45, 555–562 (2003)
3. Aktas, H., Cagman, N.: Soft sets and soft groups. Information Sciences 177, 2726–2735 (2007)
4. Acar, U., Koyuncu, F., Tanay, B.: Soft sets and soft rings. Computers and Mathematics with Applications 59, 3458–3463 (2010)
5. Jun, Y.B.: Soft BCK/BCI-algebras. Computers and Mathematics with Applications 56, 1408–1413 (2008)
6. Feng, F., Jun, Y.B., Zhao, X.: Soft semirings. Computers and Mathematics with Applications 56, 2621–2628 (2008)
7. Xiao, Z., Gong, K., Xia, S., Zou, Y.: Exclusive disjunctive soft sets. Computers and Mathematics with Applications 59, 2128–2137 (2010)
8. Herawan, T., Mat Deris, M.: A direct proof of every rough set is a soft set. In: Proceeding of the Third Asia International Conference on Modeling and Simulation, AMS 2009, Bali, Indonesia, pp. 119–124. IEEE Press, Los Alamitos (2009)
9. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9, 589–602 (2001)
10. Majumdar, P., Samanta, S.K.: Generalized fuzzy soft sets. Computers and Mathematics with Applications 59, 1425–1432 (2010)
11. Maji, P.K.: More on intuitionistic fuzzy soft sets. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS, vol. 5908, pp. 231–240. Springer, Heidelberg (2009)
12. Maji, P.K., Biswas, R., Roy, A.R.: Intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 9, 677–692 (2001)
13. Maji, P.K., Roy, A.R., Biswas, R.: On intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 12, 669–683 (2004)
14. Yang, X.B., Lin, T.Y., Yang, J., Dongjun, Y.L.A.: Combination of interval-valued fuzzy set and soft set. Computers and Mathematics with Applications 58, 521–527 (2009)
15. Jiang, Y., Tang, Y., Chen, Q., Liu, H., Tang, J.: Interval-valued intuitionistic fuzzy soft sets and their properties. Computers and Mathematics with Applications 60, 906–918 (2010)

16. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
17. Maji, P.K., Roy, A.R.: A fuzzy soft set theoretic approach to decision making problems. Journal of Computational and Applied Mathematics 203, 412–418 (2007)
18. Kong, Z., Gao, L.Q., Wang, L.F.: Comment on "A fuzzy soft set theoretic approach to decision making problems". Journal of Computational and Applied Mathematics 223, 540–542 (2009)
19. Feng, F., Jun, Y.B., Liu, X., Li, L.: An adjustable approach to fuzzy soft set based decision making. Journal of Computational and Applied Mathematics 234, 10–20 (2010)
20. Feng, F., Li, Y., Leoreanu-Fotea, V.: Application of level soft sets in decision making based on interval-valued fuzzy soft sets. Computers and Mathematics with Applications 60, 1756–1767 (2010)
21. Jiang, Y., Tang, Y., Chen, Q.: An adjustable approach to intuitionistic fuzzy soft sets based decision making. Applied Mathematical Modelling 35, 824–836 (2011)
22. Gorzalczany, M.B.: A method of inference in approximate reasoning based on interval valued fuzzy sets. Fuzzy Sets and Systems 21, 1–17 (1987)
23. Atanassov, K., Gargov, G.: Interval valued intuitionistic fuzzy sets. Fuzzy Sets and Systems 31, 343–349 (1989)

Complex-Fuzzy Adaptive Image Restoration — An Artificial-Bee-Colony-Based Learning Approach

Chunshien Li and Fengtse Chan

Laboratory of Intelligent Systems and Applications
Department of Information Management
Nantional Central University, Taiwan, ROC
jamesli@mgt.ncu.edu.tw

Abstract. A complex-fuzzy approach using complex fuzzy sets is proposed in the paper to deal with the problem of adaptive image noise cancelling. A image may be corrupted by noise, resulting in the degradation of valuable image information. Complex fuzzy set (CFS) is in contrast with traditional fuzzy set in membership description. A CFS has the membership state within the complex-valued unit disc of the complex plane. Based on the membership property of CFS, we design a complex neural fuzzy system (CNFS), so that the functional mapping ability by the CNFS can be augmented. A hybrid learning method is devised for training of the proposed CNFS, including the artificial bee colony (ABC) method and the recursive least square estimator (RLSE) algorithm. Two cases for image restoration are used to test the proposed approach. Experimental results are shown with good restoration quality.

Keywords: complex fuzzy set (CFS), complex neuro-fuzzy system (CNFS), artificial bee colony (ABC), recursive least square estimator (RLSE), image restoration.

1 Introduction

Although images can contain much useful information, image corruption by interfering signal happens in transmitting, receiving or transforming process. Interfering signal is called noise, which can be from natural environment, wireless transmission equipments, defected electronic devices, or others. Noise from interfering source may be transformed further by some unknown channel or mechanism to become another type of noise. Such a transformed noise is the interfering signal, called the attacking noise, that can attack and corrupt images into degraded quality. And, noise channel or mechanism is usually unknown and changing in real world. To solve the problem of image corruption, researchers have developed a number of different filters [1],[7],[14],[21]-[24] for noise cancelling. In this paper, we propose an artificial-intelligence-based approach to the problem of adaptive image noise cancelling (AINC). The proposed approach utilizes the theory of complex fuzzy set (CFS), the methodology of neuro-fuzzy computing, and the optimization methods of both the artificial bee colony (ABC) algorithm and the recursive least squares estimator (RLSE) algorithm. Neural networks (NN) have excellent performance in functional mapping capability, but they are lack of

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 90–99, 2011.

interpretation transparency to human and are usually viewed as black-box adaptive systems. In contrast, fuzzy inference systems (FIS) can be used to describe the human's experiences and knowledge in terms of If-Then rules. Moreover, both of NN and FIS are with the property of being universal approximator which can approximate any function to any degree of accuracy in theory. Thus, it is natural to combine them together to form a neuro-fuzzy system (NFS) [9], which takes the advantages of the learning capability by NN and the inference ability by FIS and still keep the universal approximation property. The NFS theory [8]-[10],[18] has been used in various system modeling problems and real world applications.

In this paper, we propose a complex neuro-fuzzy system (CNFS), using complex fuzzy sets, for the problem of adaptive image noise cancelling. The concept of CFS is distinct from traditional fuzzy set whose membership state is characterized in a real-valued interval between 0 and 1. In contrast, the membership state of a CFS is characterized in the complex-valued unit disc of the complex plan [6],[19]. The property of CFS [6],[17],[19] can be used to expand the capability of the function mapping by a NFS. For the training of the proposed CNFS, we devise a hybrid self-learning method, called the ABC-RLSE method, combining the ABC algorithm [12] and the RLSE algorithm [10]. The ABC algorithm is used to update the premise parameters of the CNFS and the RLSE algorithm is used to adjust the consequent parameters. The hybrid learning scheme is based on the concept of divide-and conquer, with which fast training for the CNFS can be achieved.

In Section 2, we introduce the theory of complex fuzzy set, the methodology of NFS computing, and the method of the proposed ABC-RLSE algorithm. In Section 3, the CNFS-based research architecture for the AINC study is specified. In Section 4, two cases for image restoration are given to test the proposed approach. Finally, the paper is concluded.

2 Methodology of the Proposed Approach

Based on the theory of complex fuzzy set (CFS) [6],[19]-[20] the membership state of a CFS is within a complex-valued unit disc of the complex plane. Complex fuzzy sets are different from fuzzy complex numbers [2]-[5]. The membership function of CFS is consists of an amplitude function and a phase function. For a complex fuzzy set A, the membership function $\mu_A(h)$ is defined as follows.

$$\begin{aligned}\mu_A(h) &= r_A(h)\exp\left(j\omega_A(h)\right)\\ &= \mathrm{Re}(h) + \ j\mathrm{Im}(\mu_A(h))\\ &= r_A(h)\cos\left(j\omega_A(h)\right) + jr_A(h)\sin(j\omega_A(h))\end{aligned} \tag{1}$$

where $j = \sqrt{-1}$, h is the base variable, $r_A\ (h)$ is the amplitude function and $w_A\ (h)$ is the phase function. We design a Gaussian-type complex fuzzy set as follows.

$$cGaussian(h, m, \sigma) \ = \ r_s(h, m, \sigma)\exp(jw_s(h, m, \sigma)) \tag{2a}$$

where

$$cGaussian\ (h,m,\sigma) = \exp\left(-\frac{(h-m)^2}{2\sigma^2}\right) + j\frac{-(h-m)}{2\sigma^2}\exp\left(-\frac{(h-m)^2}{2\sigma^2}\right) \quad (2b)$$

In (2a) to (2b), $\{h,m,\sigma\}$ are the base variable, mean, spread for the Gaussian type CFS.

In this paper, we use the Gaussian type complex fuzzy sets to design the proposed the complex neuro-fuzzy system (CNFS). We suppose the CNFS consists of K first-order Takagi-Sugeno (T-S) fuzzy rules, each of which has M inputs and one output, given as follows.

$$\begin{aligned}&\text{Rule } i:\ \text{IF } \left(x_1 \text{ is } A_1^i(h_1)\right) \text{ and } \left(x_2 \text{ is } A_2^i(h_2)\right) \cdots \text{and } \left(x_M \text{ is } A_1^i(h_M)\right)\\ &\qquad \text{Then } z^i = a_0^i + \textstyle\sum_{j=1}^{M} a_j^i h_j\end{aligned} \quad (3)$$

for i =1,2,...,K, where x_j is the j-th linguistic variable, h_j is the j-th input base variables, $A_j^i(h_j)$ is the j-th premise complex fuzzy set of the i-th rule, z_i is the output of the i-th rule, and $\{a_j^i,\ j$=0,1,…,$M\}$ are consequent parameters of the i-th rule. The complex fuzzy inference procedure of the CNFS is cast into six-layered neural-net structure, specified as follows. **Layer 0:** This layer receives the inputs and transmits them to the layer 1 directly. The input vector at time t is given as follows.

$$\mathrm{H}(t) = [h_1(t), h_2(t), \ldots, h_M(t)]^{\mathrm{T}} \quad (4)$$

Layer 1: The layer is called the fuzzy-set layer. Each node in the layer represents a linguistic value characterized by a complex fuzzy set. The Gaussian type of complex fuzzy set given as (2a) to (2b) is used for the design of complex fuzzy sets. **Layer 2:** This layer is for the firing-strengths of fuzzy rules. The firing strength of the i-th rule is calculated as follows.

$$\begin{aligned}\beta^i(t) &= \wedge\left(\mu_1^i(h_1(t)), \mu_2^i(h_2(t)), \cdots \mu_M^i(h_M(t))\right)\\ &= \textstyle\prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)\end{aligned} \quad (5)$$

where $\wedge$ represents the *fuzzy-and* operation, for which the product operator is used in the study; r_j^i is the amplitude of complex membership for the j-th fuzzy set of the i-*th* rule; $\omega_{A_1^i \cap \ldots \cap A_M^i}$ is the phase of the firing strength. **Layer 3:** This layer is for the normalization of the firing strengths. The normalized firing strength for the i-th rule is represented as follows.

$$\lambda^i(t) = \frac{\beta^i(t)}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \beta^i(t)} = \frac{\prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \quad (6)$$

Layer 4: The layer called the consequent layer. The normalized consequent of the i-th rule is represented as follows.

$$\begin{aligned}\xi^i(t) &= \lambda^i(t) \times z^i(t) \\ &= \lambda^i(t) \times \left(a_0^i + \sum_{j=1}^{M} a_0^i h_i(t)\right) \\ &= \frac{\prod_{j=1}^{M} r_j^i\left(h_j(t)\right) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \prod_{j=1}^{M} r_j^i\left(h_j(t)\right) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \times \left(a_0^i + \sum_{j=1}^{M} a_j^i h_j(t)\right)\end{aligned} \tag{7}$$

Layer 5: This layer is called the output layer. The normalized consequents from Layer 4 are congregated into the layer to produce the CNFS output, given as follows.

$$\begin{aligned}\xi(t) &= \sum_{i=1}^{k} \xi^i(t) = \sum_{i=1}^{k} \lambda^i(t) \times z^i(t) \\ &= \sum_{i=1}^{k} \left\{ \frac{\prod_{j=1}^{M} r_j^i(h_j(t)) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \prod_{j=1}^{M} r_j^i(h_j(t)) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \times \left(a_0^i + \sum_{j=1}^{M} a_j^i h_j(t)\right) \right\}\end{aligned} \tag{8}$$

The output of the CNFS is complex-valued, which can be expressed as follows.

$$\begin{aligned}\xi(t) &= \xi_{\mathrm{Re}}(t) + j\xi_{\mathrm{Im}}(t) \\ &= |\xi(t)| \times \exp\left(j\omega_\xi\right) \\ &= |\xi(t)| \times \cos\left(j\omega_\xi\right) + \mathrm{j}|\xi(t)| \times \sin\left(j\omega_\xi\right)\end{aligned} \tag{9}$$

where $\xi_{\mathrm{Re}}(t)$ is the real part of the output of the CNFS, and $\xi_{\mathrm{Im}}(t)$ is the imaginary part. Based on (9), the complex inference system can be viewed as a complex function, expressed as follows.

$$\xi(t) = \mathrm{F}(\mathbf{H}(t), \mathbf{W}) = \mathrm{F}_{\mathrm{RE}}(\mathbf{H}(t), \mathbf{W}) + j\mathrm{F}_{\mathrm{Im}}(\mathbf{H}(t), \mathbf{W}) \tag{10}$$

where $\mathrm{F}_{\mathrm{Re}}(.)$ is the real part of the CNFS output, $\mathrm{F}_{\mathrm{Im}}(.)$ is the imaginary part of the output, $\mathbf{H}(t)$ is the input vector to the CNFS, $\mathbf{W}$ denotes the parameter set of the CNFS, which is divided into the subset of the premise parameters and the subset of the consequent parameters, denoted as $\mathbf{W}_{\mathrm{If}}$ and $\mathbf{W}_{\mathrm{Then}}$, respectively.

$$\mathrm{W} = \mathrm{W}_{\mathrm{If}} \cup \mathrm{W}_{\mathrm{Then}} \tag{11}$$

For the training of the proposed CNFS, we devise a hybrid learning method, containing the artificial bee colony (ABC) optimization method and the recursive least squares estimator (RLSE) method. Both methods cooperate each other in hybrid way to become the ABC-RLSE learning method. The $\mathbf{W}_{\mathrm{If}}$ and $\mathbf{W}_{\mathrm{Then}}$ are updated by the ABC and RLSE, respectively.

Artificial bee colony (ABC) algorithm is a novel optimization method [11]-[13], [15]-[16] to simulate the search behavior by honey bees for nectar. For a swarm of bees, according to the nature of work, foraging bees are classified into employed bees,

onlooker bees and scout bees. Employed bees are flying to the food sources they visited previously to bring back food, and they can reveal the information of food sources by dancing while they are back at the hive. Onlooker bees are waiting on the dancing information concerning the food sources by the employed bees for making decision to select a food source. This selection is dependent on the nectar amount of each food source. Scout bees are flying randomly in the search area to find other new food sources. The nectar amount of a food source corresponds to the fitness value of the specific optimization problem, and the location of a food source represents a candidate solution to the optimization problem. Assume there are S food sources in total, where the location of the i-th food source is expressed as $\mathbf{X}_i$=[x_{i1}, x_{i2}, …, x_{iQ}] for i=1,2,…,S. In the ABC algorithm, the location of the i-th food source is updated by the following equation.

$$x_{ij}(t+1)= x_{ij}(t)+\varphi_{ij}\,(x_{ij}(t)- x_{kj}(t)) \tag{12}$$

for i=1,2,…,S and j=1,2,…,Q, where $x_{ij}(t)$ is the j-the dimensional coordinate of the i-th food source at iteration t, k is a random integer in the set of {1,2,…,S} with the constraint of $i\neq k$, and φ_{ij} in a uniform number randomly distributed in [-1, 1]. An onlooker bee goes the vicinity around X_i by the probability given below.

$$p_i = \frac{F_{fitness}(\mathbf{X}_i)}{\sum_{j=1}^{S} F_{fitness}(\mathbf{X}_j)} \tag{13}$$

where $F_{fitness}(.)$ is the fitness function. In the ABC, if the fitness of a food source is not improved further through a predetermined number of cycles, called *limit*, then that food source is assumed to be abandoned. The operation of the ABC is specified in steps. **Step1:** initialize necessary settings for the ABC. **Step 2:** send employed bees to food sources and compute their fitness values. **Step 3:** send onlooker bees to the food sources with the probability in (13) and compute their fitness values. **Step 4:** send scout bees for other new food sources. **Step 5:** update the locations of food sources and save the best so far. **Step 6:** if termination condition is met, stop; otherwise increase the iteration index and go to **Step 2** to continue the procedure.

The least squares estimation (LSE) problem can be specified with a linear model, given as follows.

$$y = \theta_1 f_1(u) + \theta_2 f_2(u) + \cdots + \theta_m f_m(u) + \varepsilon \tag{14}$$

where y is the target, u is the input to model, {$f_i(u)$, i=1,2,..,m} are known functions of u, {θ_i , i=1,2,…,m} are the unknown parameters to be estimated, and ε is the model error. Note that the parameters {θ_i, i=1,2,…,m} can be viewed as the consequent parameters of the proposed CNFS. The observed samples are collected to use as training data for the proposed CNFS. The training data (TD) is denoted as follows.

$$\mathrm{TD} = \{(u_i, y_i), i = 1,2, \dots, N\} \tag{15}$$

where (u_i, y_i) is the i-th data pair in the form of (*input, target*). With (15), the LSE model can be expressed in matrix form, $\mathbf{y}=\mathbf{A}\boldsymbol{\theta}+\boldsymbol{\varepsilon}$.

The optimal estimation for $\boldsymbol{\theta}=[\theta_1, \theta_2,\ldots\theta_m]^{\mathrm{T}}$ can be calculated using the following recursive least squares estimator (RLSE) equations recursively.

$$\mathbf{P}_{k+1} = \mathbf{P}_k - \frac{\mathbf{P}_k \mathbf{b}_{k+1} \mathbf{b}_{k+1}^T \mathbf{P}_k}{1 + \mathbf{b}_{k+1}^T \mathbf{P}_k \mathbf{b}_{k+1}} \tag{16a}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{P}_{k+1} \mathbf{b}_{k+1} (y_{k+1} - \mathbf{b}_{k+1}^T \boldsymbol{\theta}_k) \tag{16b}$$

for k = 0,1,2,…,(*N-1*), where $\boldsymbol{\theta}_k$ is the estimator at the k-th iteration and $[\mathbf{b}_k^T, y_k]$ is the k-th row of [**A**, **y**]. To implement the RLSE in (16a) and (16b), we initialize $\boldsymbol{\theta}_0$ with the zero vector and $\mathbf{P}_0=\alpha\mathbf{I}$, where α is a large positive number and **I** is the identity matrix.

3 Strategy of Image Noise Cancelling by the Proposed Approach

Based on the famous adaptive noise cancelling (ANC) method [24], we propose the complex fuzzy approach to the problem of adaptive image noise cancelling (AINC). The strategy for the AINC is given in diagrammatic way, shown in Fig. 1, where the CNFS given previously is used as an adaptive filter, and a median filter is involved in the diagram as well. The motivation of the usage of the median filter is that we try to get information of noise in any way from the difference of the median filter output and the corrupted image pixels, so that the CNFS filter can be trained to mimic the dynamic behavior of the noise channel. The proposed AINC procedure is given as follows.

Step 1. Initialize the CNFS design and the settings for ABC-RLSE learning method.
Step 2. Calculate the average gray level of the corrupted image, as follows.

$$\mu_Z = \frac{\sum_{i=1}^{P} \sum_{j=1}^{Q} z(i,j)}{P \times Q} \tag{17}$$

where $z(i,j)$ indicates the gray level of the image; $P \times Q$ represents the dimensions of the image. Change the corrupted image to its zero-mean version, using (17).

Step 3. Compute the difference between the median filter output and the corrupted image pixel in gray level. The difference becomes input to the CNFS.
Step 4. Calculate the CNFS output pixel by pixel, with which the cost is computed as follows.

$$\mathrm{MSE} = \frac{1}{N} \sum_{\mathrm{t}=1}^{N} \big(G(t) - \xi(t)\big)^2 \tag{18}$$

Step 5. Apply the ABC-RLSE learning method to the CNFS filter, where the subset of the premise parameters is updated by the ABC and the subset of the consequent parameters is updated by the RLSE. Repeat Step 4 until stop condition is met.
Step 6. Perform image restoration.

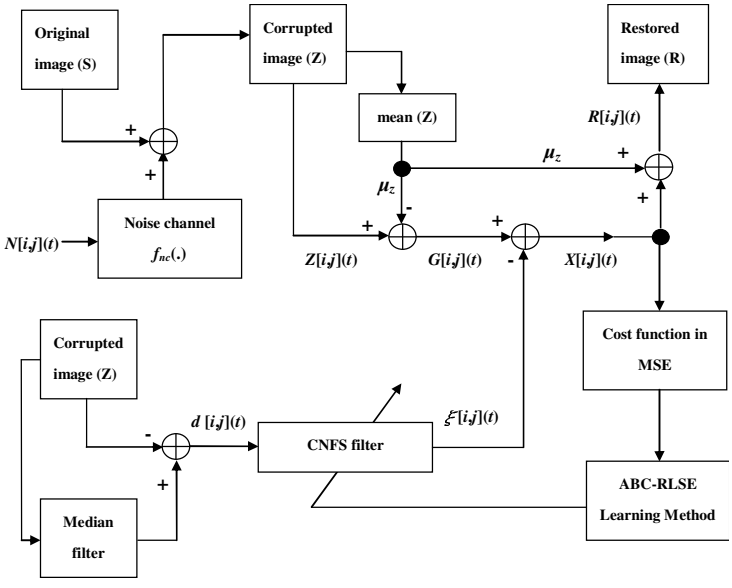


Fig. 1. Image noise cancelling by the proposed CNFS approach

4 Experimentation for the Proposed Approach

There are nine T-S fuzzy rules designed for the CNFS, including 12 premise parameters and 27 consequent parameters. The cost function in MSE is designed, as given in (**18**). And operator windows of mean and median filter are 3 × 3 in the experimentation. The peak signal to noise ratio (PSNR), $\mathrm{PSNR} = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right)$, is used as the performance index for restoration. The input vector to the CNFS is given as $\mathbf{H}(t)=[\, d(k\text{-}1), d(k)]^{\mathrm{T}}$, where $d(k)$ is the difference between the median filter output and the corrupted image pixel gray level at time k. The settings for the ABC-RLSE learning algorithm are shown in Table 1 and the predetermined *limit* for the ABC is set as 100. The "football" image is used in the experimentation, which has image resolution of 256 × 320. The first 600 pixels are involved to the training of the CNFS filter. To test the proposed approach, we use two different nonlinear transfer functions for the noise channel, $f1_{\mathrm{nc}}(.)$ and $f2_{\mathrm{nc}}(.)$, given below.

Table 1. Settings for the ABC-RLSE hybrid learning method

ABC settings		RLSE settings	
Number of premise parameters	12	Number of consequent parameters	27
Bee swarm size	10	**θ**	27×1
Employed bees	5	$\mathbf{P}_0$	A**i**
Onlooker bees	5	α	10^8
Scout bees	1	**I**	27×27identity matrix

$$f1_{\mathrm{nc}}(N) = 10 \times N + 5 \times \cos(2 \times \pi \times N + 1.5) - 10 \tag{19}$$

$$f2_{\mathrm{nc}}(N) = 10 \times N^2 - 5 \times \pi \tag{20}$$

where N is Gaussian random. The original and corrupted images are shown in Figs. 2 and 3, respectively. The proposed approach is compared to the median filtering approach and the mean filtering approach. For the case of noise channel with $f1_{\mathrm{nc}}(.)$, the restored image by the proposed approach is shown in Figure 4, and the curve of the corresponding pixel levels (shown first 1000 pixels) is given in Figure 5. Similarly, for the case of noise channel with $f2_{\mathrm{nc}}(.)$, the results are shown in Figures 6 and 7. The performance comparison in PSNR is given in Table 2.

Table 2. Performance comparison in PSNR

Noise channel	Mean filtering	Median filtering	Proposed approach
$f1_{\mathrm{nc}}(.)$	25.491	25.362	29.352
$f2_{\mathrm{nc}}(.)$	26.202	26.977	29.575

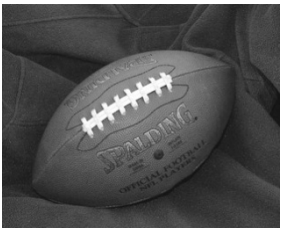


Fig. 2. Original "football" image

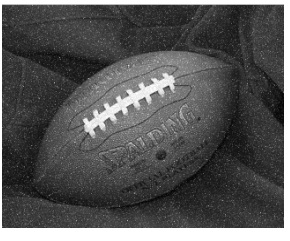


Fig. 3. Corrupted image

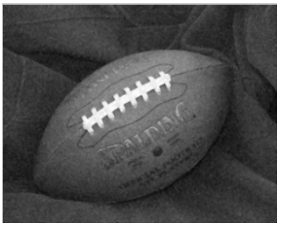


Fig. 4. Restored image by the proposed approach, (the $f1_{\mathrm{nc}}(.)$ case)

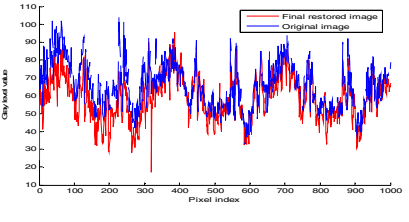


Fig. 5. First 1000 pixel-level curve of the restored image by the proposed approach, compared to the original image, (the $f1_{\mathrm{nc}}(.)$ case)

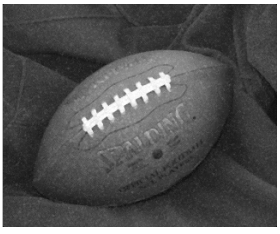


Fig. 6. Restored image by the proposed approach, (the $f2_{nc}(.)$ case)

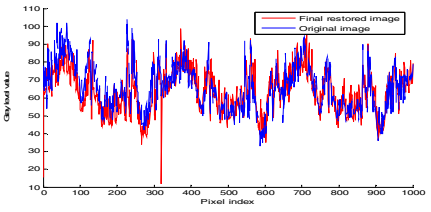


Fig. 7. First 1000 pixel-level curve of the restored image by the proposed approach, compared to the original image, (the $f2_{nc}(.)$ case)

5 Discussion and Conclusion

We have presented the proposed CNFS approach to the problem of adaptive image noise cancelling, using the theory of complex fuzzy set, the neuro-fuzzy computing rationale, and the hybrid learning method which includes the bee-swarm-intelligence optimization method and recursive least squares estimator method for the training of the CNFS. For performance comparison, the proposed approach has been compared to the mean filtering approach and the median filtering approach, as shown in Table 2, where the proposed approach outperforms the compares approaches. With the ABC-RLSE hybrid learning algorithm, the parameters of the CNFS filter is separated into two smaller subsets $\mathbf{W}_{\text{If}}$ and $\mathbf{W}_{\text{Then}}$, based on the concept of divide-and-conquer. The parameters of $\mathbf{W}_{\text{If}}$ and $\mathbf{W}_{\text{Then}}$ are updated by the ABC and the RLSE, respectively, so that the optimal solution to the parameters of the CNFS can be found easier and faster. According to experimental results, the proposed approach has shown good performance for adaptive image restoration.

References

1. Brownrigg, D.R.K.: The weighted median filter. Commun. Assoc. Comput. Mach. (ACM) 27, 807–818 (1984)
2. Buckley, J.J., Hu, Y.: Fuzzy complex analysis I: Differentiation. Fuzzy Sets and Systems 41, 269–284 (1991)
3. Buckley, J.J.: Fuzzy complex analysis II: Integration. Fuzzy Sets and Systems 49, 171–179 (1992)
4. Buckley, J.J.: Fuzzy complex numbers. Fuzzy Sets and Systems 33, 333–345 (1989)
5. Castro, J.L.: Fuzzy logic controllers are universal approximators. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 25, 629–635 (1995)
6. Dick, S.: Toward complex fuzzy logic. IEEE Transactions on Fuzzy Systems 13, 405–414 (2005)
7. Etter, W., Moschytz, G.S.: Noise reduction by noise-adaptive spectral magnitude expansion. J. Audio Engineering Society 42, 341–349 (1994)
8. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)

9. Jang, S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–685 (1993)
10. Jang, J.S.R., Sum, C.T., Mizutani, E.: Neuro-fuzzy and soft computing. Prentice-Hall, Englewood Cliffs (1997)
11. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. J. Global Optimization 39, 171–459 (2007)
12. Karaboga, D., Basturk, B.: Artificial bee colony algorithm on training artificial neural networks. Signal Processing and Communications Applications, 1–4 (2007)
13. Karaboga, D., Basturk, B.: Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 789–798. Springer, Heidelberg (2007)
14. Ko, S.J., Lee, Y.H.: Center weighted median filters and their applications to image enhancement. IEEE Transactions on Systems, Circuits System 38, 984–993 (1991)
15. Mala, D.J., Mohan, V.: ABC Tester - Artificial bee colony based software test suite optimization approach. International Journal of Software Engineering, 15–43 (2009)
16. Ming, L.D., Yi, C.B.: Artificial bee colony algorithm for scheduling a single batch processing machine with non-identical job sizes. J. Sichuan University 46, 657–662 (2009)
17. Moses, D., Degani, O., Teodorescu, H.N., Friedman, M., Kandel, A.: Linguistic coordinate transformations for complex fuzzy sets. Fuzzy Systems Conference Proceedings 3, 1340–1345 (1999)
18. Mousavi, S.J., Ponnambalam, K., Karray, F.: Inferring operating rules for reservoir operations using fuzzy regression and ANFIS. Fuzzy Sets and Systems 158, 1064–1082 (2007)
19. Ramot, D., Milo, R., Friedman, M., Kandel, A.: Complex fuzzy sets. IEEE Transactions on Fuzzy Systems 10, 171–186 (2002)
20. Ramot, D., Milo, R., Friedman, M., Kandel, A.: Complex fuzzy logic. IEEE Transactions on Fuzzy Systems 11, 450–461 (2003)
21. Rosenfeld, A., Kak, A.C.: Digital picture processing. Academic Press, New York (1982)
22. Russo, F.: Noise removal from image data using recursive neurofuzzy filters. IEEE Transactions on Fuzzy Systems 49, 307–314 (2000)
23. Saudia, S., Varghese, J., Allaperumal, K.N., Mathew, S.P., Robin, A.J., Kavitha, S.: Salt & pepper impulse detection and median based regularization using adaptive median filter. In: IEEE Region 10 Conference on Innovative Technologies for Societal Transformation, pp. 1–6 (2008)
24. Widrow, B., Glover, J.R., McCool, J.M.: Adaptive noise canceling: Principles and application. Proceedings of IEEE 63, 1692–1730 (1975)

Rule Extraction for Support Vector Machine Using Input Space Expansion

Prasan Pitiranggon, Nunthika Benjathepanun, Somsri Banditvilai, and Veera Boonjing

Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Rd., Ladkrabang, Bangkok 10520, Thailand
prasan.pitiranggon@hotmail.com, kbnunthi@kmitl.ac.th, kbsomsri@kmitl.ac.th, kbveera@kmitl.ac.th
http://www.kmitl.ac.th

Abstract. Fuzzy Rule-Based System (FRB) in the form of human comprehensible IF-THEN rules can be extracted from Support Vector Machine (SVM) which is regarded as a black-boxed system. We first prove that SVM decision network and the zero-ordered Sugeno FRB type of the Adaptive Network Fuzzy Inference System (ANFIS) are equivalent indicating that SVM's decision can actually be represented by fuzzy IF-THEN rules. We then propose a rule extraction method based on kernel function firing strength and unbounded support vector space expansion. An advantage of our method is the guarantee that the number of final fuzzy IF-THEN rules is equal or less than the number of support vectors in SVM, and it may reveal human comprehensible patterns. We compare our method against SVM using popular benchmark data sets, and the results are comparable.

Keywords: Rule extraction, fuzzy IF-THEN rules, Support Vector Machine, pattern classification.

1 Introduction

Support Vector Machine (SVM) is a great tool to approximate functions, recognize patterns, or predict outcomes [13], [17]. Despite its great performance, it suffers from its black-boxed characteristics [2], [4]. To make it white-boxed, rule extraction is needed [12]. A limited number of studies of rule extraction from SVM have been conducted to obtain more understandable rules in order to explain how a decision was made. Techniques specifically intended as SVM rule extraction techniques are based on translucency and scope. Translucency can be either pedagogical or decompositional, and scope can be either classification or regression. Pedagogical techniques are those that try to relate inputs with outputs without making use of system structure, but decompositional techniques do make use of structure of the system. Classification techniques are the ones trying to differentiate input patterns, but regression techniques are trying to approximate function values.

Previous studies on rule extraction techniques for SVM using decompositional techniques are SVM + Prototype [15] which uses input clustering to obtain

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 100–109, 2011.

prototype vectors and geometrical formulas to obtain ranges for IF-THEN rules, tree related method [3] which uses a tree technique on support vectors to obtain IF-THEN rules, and cubes and separating hyperplane related [6] which uses cubes extending from separating hyperplane and can be used only in linear SVM, while the ones using pedagogical techniques are Iter [8] which uses randomly generated vectors in input space to cover entire input vectors and Minerva [9] which is similar to Iter but uses Sequential Covering method additionally.

None of the previous studies uses kernel function strength in rule extraction. Our study makes use of the kernel function strength similar to the way SVM makes decisions to extract rules; therefore, the decision rules obtained are closer to SVM decisions. Moreover, our technique guarantees that the number of final rules is less than the number of support vectors obtained by SVM. The technique used in our study is considered decompositional in translucency and classification in scope. Our technique looks for unbounded support vectors, which are the data points used as base locations to define separating hyperplane, to build trained SVM decision network to classify testing data. We can prove that SVM and a type of FRB, called Adaptive Network Fuzzy Inference System (ANFIS), are equivalent which means fuzzy IF-THEN rules can represent SVM decisions without loss of functionality. We also compare our method with SVM classifier using popular benchmark data sets as a validation.

2 Method

2.1 Finding Unbounded Support Vectors

We go through standard procedure of SVM using input data and Gaussian kernel to get unbounded support vectors [13].

2.2 Constructing Trained SVM Decision Network

We can use the unbounded support vectors obtained in the previous step to construct a trained SVM decision network. Graphical representation of trained SVM decision network, which can be used to classify input data, is shown in Fig. 1.

2.3 Proof of Functional Equivalence of SVM and FRB

The purpose of this part is to formally prove that SVM decision network is equivalent to fuzzy IF-THEN rules which means SVM decision can be represented by fuzzy IF-THEN rules without loss in functionality.

Fuzzy IF-THEN Rules and Fuzzy Inference System Functions. There are two well-known types of fuzzy inference method. Mamdani's fuzzy inference method and Takagi-Sugeno (TS) method [11]. TS method can be further divided into zero-ordered and first-ordered type. The zero-ordered type contains only

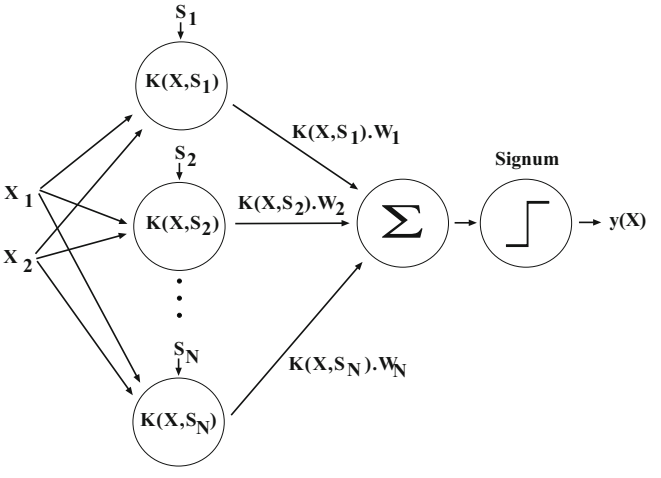


Fig. 1. A graphical representation of an SVM decision network is shown, where the leftmost X_i is input vector, S_i is unbounded support vector, $K(X, S_i)$ is Gaussian kernel function where X is each input vector, W_i is weight, and $y(X)$ is output decision

constant in its consequent, but the first-ordered type contains linear equation with variables from the antecedent part.

There is a layered network system called Adaptive Network-Based Fuzzy Inference System (ANFIS) [11] which can be made functionally equivalent to Artificial Neural Networks (ANN). ANFIS is based on Adaptive Network which contains layers of functional nodes with connectors; square nodes are dynamic nodes which depend on node parameters, and circle nodes are fixed nodes which have empty set of parameters (Fig. 2). We can obtain fuzzy IF-THEN rules from the ANFIS which is derived from ANN [10]. In this study, instead of making ANFIS equivalent to ANN, we make the ANFIS equivalent to SVM.

Fuzzy inference system is composed of a set of fuzzy IF-THEN rules, a database containing membership functions of linguistic labels, and an inference mechanism called fuzzy reasoning. Only zero-ordered TS FRB will be shown to be equivalent to SVM using the following example model as an illustration.

Suppose we have a rule base consisting of two fuzzy IF-THEN rules of TS type:

Rule 1: If x_1 is A_1 and x_2 is B_1 then $f_1 = a_1x_1 + b_1x_2 + c_1$
Rule 2: If x_1 is A_2 and x_2 is B_2 then $f_1 = a_2x_1 + b_2x_2 + c_2$

then the fuzzy reasoning mechanism can be illustrated in Fig. 2, where the firing strength of i^{th} rule is obtained as the T-norm (usually minimum or multiplication operator) of the membership values on the premise part. In our case, we only use multiplication operator in T-norm step. Strength after T-norm with multiplication operator is:

$$M_i = \mu_{A_i}(X_1)\mu_{B_i}(X_2) \ . \tag{1}$$

Note that overall output can be chosen as the weighted sum of each rule's output:

$$f(X) = \sum_{i=1}^{R} M_i.w_i \tag{2}$$

where R is the number of fuzzy IF-THEN rules. And the decision equation is:

$$y(X) = sign(f(X) + b) \ . \tag{3}$$

Required Conditions for Functional Equivalence. The functional equivalence between a trained SVM decision network (Fig. 1) and ANFIS (Fig. 2) can be established if the following are true:

- The number of input patterns is equal to the number of fuzzy IF-THEN rules.
- The output of each fuzzy IF-THEN rule is composed of a constant.
- The membership functions within each rule are chosen as Gaussian functions with the same variance.
- The T-norm operator used to compute each rule's firing strength is multiplication.
- Both the SVM decision network and the fuzzy inference system under consideration use the weighted sum method to derive their overall outputs.

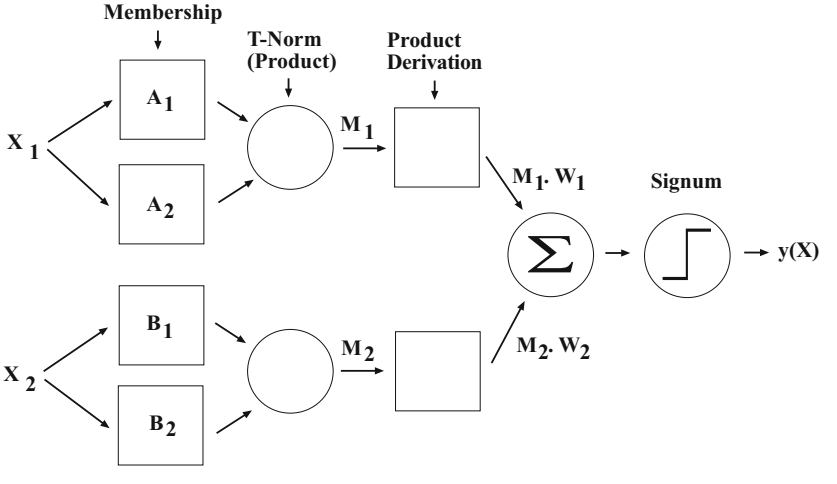


Fig. 2. An example of ANFIS equivalent to SVM, which uses Gaussian kernel function with two input vectors

Functions in SVM decision network. Let P be a set of functions:

$P = \{f_1, f_2, f_3, f_4\}$.

Let X be a set of input vectors:

$X = \{X_1, X_2, X_3, \ldots, X_n\}$ where n is the total number of input vectors.

Let S be a set of unbounded support vectors:

$S = \{S_1, S_2, S_3, \ldots, S_N\}$ where N is the total number of unbounded support vectors.
Let $f_1(x, y)$ be Gaussian kernel function; we obtain

$$f_1(X, S_i) = exp\left[\frac{-(\parallel X - S_i \parallel)^2)}{\sigma^2}\right], \quad i = 1, \ldots, N \; . \tag{4}$$

Let $f_2(x, y)$ be multiplication function; we obtain

$$f_2(f_1(X, S_i), w_i) = w_i . f_1(X, S_i) \; . \tag{5}$$

Let $f_3(x) = \sum_{i=1}^{N} x_i$, we obtain

$$f_3(X) = \sum_{i=1}^{N} f_{2_i}(X) \; . \tag{6}$$

Let $f_4(x) = sign(x + a)$, we obtain

$$f_4 = sign(f_3 + a) \; . \tag{7}$$

Let y be the output; we obtain

$$y = f_4(f_3(f_2(f_1(X, S_i)))) = f_4 \circ f_3 \circ f_2 \circ f_1(X, S_i) \; . \tag{8}$$

Functions in ANFIS. Let Q be a set of functions in ANFIS:

$Q = \{g_1, g_2, g_3, g_4\}$.

Let X be a set of input vectors:

$X = \{X_1, X_2, X_3, \ldots, X_n\}$ where n is the total number of input vectors.

Let S be a set of unbounded support vectors:

$S = \{S_1, S_2, S_3, \ldots, S_N\}$ where N is the total number of unbounded support vectors.
Let $g_1(X, S_i)$ be Gaussian membership function with T-norm operation

$$\mu_{A_1}(x_1) = exp\left[\frac{-(x_1 - c_{A_1})^2}{\sigma_1^2}\right] \tag{9}$$

where $\mu_{A_1}(x_1)$ is membership function

$$M_i = \mu_{A_i}(x_1)\mu_{B_i}(x_2) \tag{10}$$

where M_i is result of T-norm operation at i.
We obtain

$$g_1(X, S_i) = exp\left[\frac{-(\parallel X - S_i \parallel)^2}{\sigma^2}\right], \quad i = 1, \ldots, N \; . \tag{11}$$

Let $g_2(x, y)$ be multiplication function; we obtain

$$g_2(g_1(X, S_i), w_i) = w_i.g_1(X, S_i) \; . \tag{12}$$

Let $g_3(x) = \sum_{i=1}^{N} x_i$, we obtain

$$g_3(X) = \sum_{i=1}^{N} g_{2_i}(X) \; . \tag{13}$$

Let $g_4(x) = sign(x + a)$, we obtain

$$g_4 = sign(g_3 + a) \; . \tag{14}$$

Let z be the output; we obtain

$$z = g_4(g_3(g_2(g_1(X, S_i)))) = g_4 \circ g_3 \circ g_2 \circ g_1(X, S_i) \; . \tag{15}$$

The Proof of the Equivalence of SVM and ANFIS

Proof. Since $f1 = g1, f2 = g2, f3 = g3$, and $f4 = g4$, then $y = z$, and X and Si are the same input in both systems; therefore, the two systems are functionally equivalent.

2.4 Rules Extraction Based on Firing Strength

The purpose of this step is to extract rules based on the strongest firing signals associated with unbounded support vectors in high dimensional space.

In Fig. 3, all input patterns are entered into system one at a time. Gaussian kernel function as a membership function is calculated between current input and each of the support vectors with the class label we want to identify ignoring the support vectors of the other class, and the highest value is considered the strongest signal which will be the only one fired, and the rest will be ignored. The fired row then stores cumulative min and max value which will be replaced by new min or new max if it occurs. After all input patterns have been entered, min and max values in each row will be used as a range in conditional of each IF-THEN rule.

Schematic diagram of ANFIS implementing rule generation from unbounded support vectors found from previous step is shown in Fig. 3; X_i is input vector, and A_i is Gaussian kernel function of unbounded support vector and input vector.

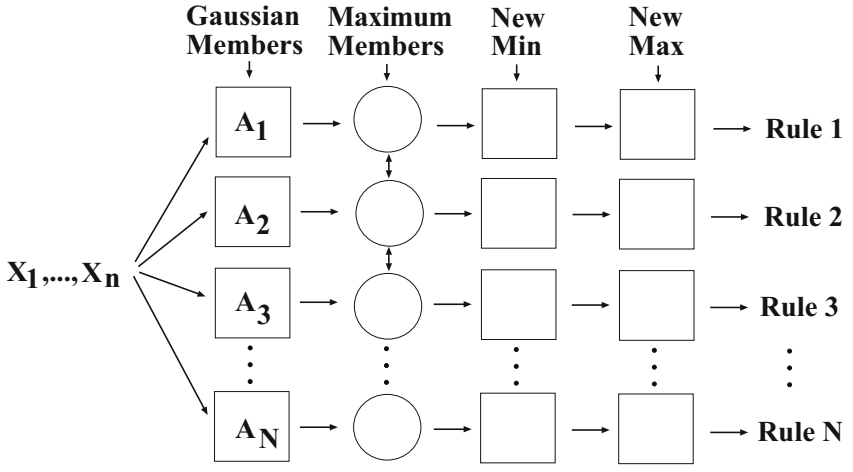


Fig. 3. Rule extraction algorithm based on kernel firing strength

The final min and max values of each row are used as a range of the newly generated IF-THEN rules. The range conditional IF-THEN statements are in the form:
Rule 1: If $(x_{11} > a_{11-}$ AND $x_{11} < a_{11+})$ AND $(x_{12} > a_{12-}\ AND\ x_{12} < a_{12+})$ AND $\ldots (x_{1n} > a_{1n-}\ AND\ x_{1n} < a_{1n+})$ THEN $y = v_1$
Rule 2: If $(x_{21} > a_{21-}$ AND $x_{21} < a_{21+})$ AND $(x_{22} > a_{22-}\ AND\ x_{22} < a_{22+})$ AND $\ldots (x_{2n} > a_{2n-}\ AND\ x_{2n} < a_{2n+})$ THEN $y = v_2$
Rule 3: If $(x_{31} > a_{31-}$ AND $x_{31} < a_{31+})$ AND $(x_{32} > a_{32-}\ AND\ x_{32} < a_{32+})$ AND $\ldots (x_{3n} > a_{3n-}\ AND\ x_{3n} < a_{3n+})$ THEN $y = v_3$
.
.
.
Rule N: If $(x_{N1} > a_{N1-}$ AND $x_{N1} < a_{N1+})$ AND $(x_{N2} > a_{N2-}\ AND\ x_{N2} < a_{N2+})$ AND $\ldots (x_{Nn} > a_{Nn-}\ AND\ x_{Nn} < a_{Nn+})$ THEN $y = v_N$
where a_{ij-} are lower range values (cumulative min) and a_{ij+} are upper range values (cumulative max) in $\Re$. N is the total number of unbounded support vectors, and n is the dimension of input vectors.

2.5 Refining Rules by Unbounded Support Vector Space Expansion

The purpose of this step is to reduce generated rules and refine rule extraction in low dimensional space. We can combine many range conditional IF-THEN statements from previous step together as long as it does not cause misclassification. Algorithms pseudo code for input space expansion is:

[Pre-loop condition: Total number of IF-THEN rules equal to total number of support vectors]
FOR i = 1 TO N

[N = total number of generated rules]
IF rule i was eliminated THEN NEXT i
DO WHILE (no class overlap from another class) or (maximum value or minimum value of the input data set reached)
Expand ranges of IF-THEN conditional at i by a small value (e.g., less than 10% of min value of an attribute)
IF there is class overlap GOTO END WHILE
END IF
END WHILE
END IF
NEXT i
FOR i = 1 TO N
DO WHILE (there are still rules to merge for this i)
IF two ranges coincide then merge the two rules by retaining the larger ranges
END IF
END WHILE
NEXT i
[Post-loop condition: Number of IF-THEN rules are the same or less than rules in pre-condition]

We can use set membership symbol in place of greater than and less than signs as our final form of rules.
IF $x_{11} \in [a_{11-}, a_{11+}]$ AND $x_{12} \in [a_{12-}, a_{12+}]$ AND $\ldots x_{1n} \in [a_{1n-}, a_{1n+}]$ THEN $y = v_1$

3 Experimental Results

We perform classification using both SVM and our fuzzy IF-THEN rules on six benchmark data sets which can be downloaded from UCI Machine Learning Repository at http://www.ics.uci.edu. The six data sets are Iris [5], Wine [1], Wisconsin Breast Cancer [18], Habermans Survival Data [7], Ionosphere [16], and Spect Heart [14]. These data sets are chosen to represent different number of instances, number of classes, input data types, and number of attributes. Data Characteristics are data set name (Data Set), number of instances (N), number of classes (Class), data type (Type) which can be real (R), integer (I), or binary (B), and number of attributes (Attribute). Each data set is randomly split into ten parts. In rotation, nine parts are the training data and the remaining part is the testing data in a ten-fold cross validation technique except the Spect Heart which contains its own training and testing data separately. Procedures in the method section are then applied to these data.

Results from each data set are average number of unsigned support vectors (Avg USV), average percent error (Avg %Error) from SVM, average number of rules (Avg Rules) from our method, and average percent error from our method, and the classification results are shown in Table 1. SVM performs classification

Table 1. Comparison of Errors from SVM and Rules

Data Characteristics					SVM		Rules	
Data Set	N[a]	Class	Type	Attribute	Avg USV	Avg %Error	Avg Rules	Avg %Error
Iris	150	3	R	4	37.07	2.89	7.27	6.22
Wine	178	3	R	13	157.57	14.04	52.53	8.99
Wisconsin	699	2	I	10	300.6	5.27	67.5	4.83
Haberman	306	2	I	4	105.7	29.08	63.7	46.41
Ionosphere	351	2	I, R	34	278.0	37.04	165.4	6.84
Spect Heart	187	2	B	22	73.0	10.16	20.0	36.36

[a] Number of instances.

with less error in Iris, Haberman, and Spect Heart data sets than our method, but our method performs better in Wine, Wisconsin Breast Cancer, and Ionosphere data sets. Average number of rules in each data set is lower than number of unsigned support vectors as claimed by our method.

4 Conclusion

The proposed method is shown to be a good alternative method for rule extraction from SVM and has an advantage over the decision method of SVM by revealing reasons behind the decision. And this makes it more attractive to be used in classification or prediction whenever we want to have insight into the way classification decision is made plus the fuzzy IF-THEN rules obtained can be easily incorporated into computer program.

The results of our experiments have shown that our method can outperform SVM decisions in some data sets, but most percent errors of the two methods are not far apart. It can be stated that the results of the errors from the two methods are comparable. Another advantage of our method compared to others is the guarantee that the number of rules in the final set will not exceed the number of support vectors.

One of the suggestions for future study would be to use clustering algorithm in high noise data sets. In data sets with high noise, performance of IF-THEN rules by our algorithm may not perform well in classification. Also if there are large number of input data, scalability will suffer. K means clustering may be used in these cases to handle noisy data and also help scalability. After k means clustering is run, our algorithm can be implemented to obtain IF-THEN rules from SVM. Another suggestion for future study would be a modification of our method to handle data sets with a categorical data type.

References

1. Aeberhard S., Coomans D., de Vel, O.: The Classification Performance of RDA. Tech. Rep., no. 92-01 (1992)
2. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 8(6), 373–389 (1995)

3. Barakat, N., Diederich, J.: Eclectic Rule-Extraction from Support Vector Machines. International Journal of Computational Intelligence 2(1), 59–62 (2005)
4. Diederich, J. (ed.): Rule Extraction from Support Vector Machines. SCI, vol. 80, pp. 3–30. Springer, Heidelberg (2008)
5. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annual Eugenics 7(Part II), 179–188 (1936)
6. Fung, G., Sandilya, S., Rao, R.B.: Rule extraction from linear support vector machines. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 32–40 (2005)
7. Haberman, S.J.: Generalized Residuals for Log-Linear Models. In: Proceedings of the 9th International Biometrics Conference, Boston, pp. 104–122 (1976)
8. Huysmans, J., Baesens, B., Vanthienen, J.: ITER: An algorithm for predictive regression rule extraction. In: Tjoa, A.M., Trujillo, J. (eds.) DaWaK 2006. LNCS, vol. 4081, pp. 270–279. Springer, Heidelberg (2006)
9. Huysmans, J., Setiono, R., Baesens, B., Vanthienen, J.: Minerva: Sequential Covering for Rule Extraction. IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics 38(2), 299–309 (2008)
10. Jang, J.S.R., Sun, C.T.: Functional equivalence between radial basis function networks and fuzzy inference systems. IEEE Trans. Neural Networks 4, 156–158 (1992)
11. Jang, J.S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing, pp. 333–342. Prentice Hall International, Englewood Cliffs (1997)
12. Kolman, E., Margaliot, M.: Are artificial neural networks white boxes? IEEE Trans. Neural Networks 16(4), 844–852 (2005)
13. Kumar, S.: Neural Networks: A Classroom Approach, International Edition, pp. 273–304. McGraw-Hill, New York (2005)
14. Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M., Goodenday, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23(2), 149–169 (2001)
15. Nunez, H., Angulo, C., Catala, A.: Rule Extraction Based on Support and Prototype Vectors. SCI, vol. 80, pp. 109–134. Springer, Heidelberg (2008)
16. Sigillito, V.J., Wing, S.P., Hutton, L.V., Baker, K.B.: Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest 10, 262–266 (1989)
17. Vapnik, V.N.: Statistical Learning Theory, pp. 375–520. John Wiley & Sons, Chichester (1998)
18. Wolberg, W.H., Mangasarian, O.L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences 87, 9193–9196 (1990)

Uniform RECA Transformations in Rough Extended Clustering Framework

Dariusz Małyszko and Jarosław Stepaniuk

Department of Computer Science
Biaystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
{d.malyszko,j.stepaniuk}@pb.edu.pl

Abstract. In the paper, in the Rough Extended Framework, a new generalization of the concept of the rough transformation has been presented. The introduced solution seems to present promising area of data analysis, particulary suited in the area of image properties analysis. The uniform *RECA* transformation as a generalization of clustering approaches contains three standard rough transformations - standard k-means transformation, fuzzy k-means transformation and *EM* k-means transformation. The concept of the *RECA* transformations has been illustrated with its application in the procedure of calculation of the entropy of the RECA transformation paths. In this way, uniform RECA transformations give both the theoretical ground for three most prominent data clustering schemes and at the same time present starting point in the new data analysis methodology based upon the new introduced concept of RECA paths.

1 Introduction

Data clustering routines have emerged as most prominent and important data analysis methods that are primarily applied in unsupervised learning and classification problems. Most often data clustering presents descriptive data grouping that identifies homogenous groups of data objects on the basis of the feature attributes assigned to clustered data objects. In this context, a cluster is considered as a collection of similar objects according to predefined criteria and dissimilar to the objects belonging to other clusters.

Data analysis based on the fuzzy sets depends primarily on the assumption, stating that data objects may belong in some degree not only to one concept or class but may partially participate in other classes.

Rough set theory [6], [9] on the other hand assigns objects to class lower and upper approximations on the base of complete certainty about object belongingness to the class lower approximation and on the determination of the possible belongingness to the class upper approximation. In this way, rough set based data analysis approaches seem to be advantageous during managing uncertainty and extracting knowledge form uncertain or incomplete data sources.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 110–119, 2011.

Probabilistic approaches have been developed in several rough set settings, including decision-theoretic analysis, variable precision analysis, and information-theoretic analysis. Most often, probabilistic data interpretation depends upon rough membership functions and rough inclusion functions.

Rough Extended (Entropy) Framework presents extensively developed hybrid method of data analysis that combines the best elements from all the above mentioned theories.

In the paper, a new generalization of the concept of the rough transformation has been presented as uniform *RECA* transformations. The introduced solution presents emerging promising area of data analysis, particulary suited in the area of image properties analysis. The uniform *RECA* transformation as a generalization of clustering approaches contains standard k-means transformation, fuzzy k-means transformation and EM k-means transformation. Additionally, the uniform RECA transformations give a medium for the introduction of the RECA transformation paths and the entropy of RECA transformation paths. This latter notion seems to be potentially robust approach for detailed data properties analysis.

This paper has been structured in the following way. In Section 2 the introductory information about rough sets in the context of developed Rough Extended Clustering Framework has been presented. In Section 3 the concepts of uniform *RECA* transformations have been introduced and further detailed described in Section 4. The algorithmic material has been followed by concluding remarks.

2 Rough Extended Clustering Framework

2.1 Rough Set Theory Essentials

An information system is a pair *(U,A)* where U represents a non-empty finite set called the universe and A a non-empty finite set of attributes [6]. Let $B \subseteq A$ and $X \subseteq U$. Taking into account these two sets, it is possible to approximate the set X making only the use of the information contained in B by the process of construction of the lower and upper approximations of X and further to express numerically the roughness $R(AS_B, X)$ of a set X with respect to B by assignment

$$R(AS_B, X) = 1 - \frac{card(LOW(AS_B, X))}{card(UPP(AS_B, X))}. \tag{1}$$

The value of the roughness of the set X equal 0 means that X is crisp with respect to B, and conversely if $R(AS_B, X) > 0$ then X is rough (i.e., X is vague with respect to B). Detailed information on rough set theory is provided in [7,9]. During the last decades, the rough set theory has been developed, examined and extended independently in many innovative fuzzy, probabilistic [10] and hybrid frameworks [8], [9] that combine different data analysis approaches.

Rough **E**xtended **F**ramework presents data analysis system based upon merging new data analysis tools into rough set theory, see [5], [1], [2], [4]. In REF framework data object properties and structures are analyzed by means of their

relation to the selected set of data objects from the data space. This reference set of data objects performs as the set of thresholds or the set of cluster centers. In this context, Rough Extended Framework basically consists of two interrelated approaches, namely Rough Extended Thresholding Framework (T-REF) [3] and Rough Extended Clustering Framework (C-REF). Each of these approaches gives way development and calculation of rough measures. Rough measures based upon entropy notion are further referred to as rough entropy measures.

The subject of this paper consists in the application of the clustering notions of C-REF platform in the form of uniform RECA transformations.

2.2 General Concepts of Rough Extended Clustering Framework

In general Rough Extended Clustering Framework data object are analyzed by means of their relation to the selected number of cluster centers. Cluster centers are regarded as representatives of the clusters. The main assumption made during *REF* based analysis consists on the remark that the way data objects are distributed in the clusters determines internal data structure. In the process of the inspection of the data assignment patterns in different parametric settings it is possible to reveal or describe properly data properties.

Data objects are assigned to lower and upper approximation on the base of the criteria given in Table 1 for difference based thresholds and in Table 2 for distance based thresholds. The calculation of the RECA-S sets gives, for example the lower and upper approximations, gives the possibility and theoretical background for development of robust clustering schemes. After the measures for

Table 1. Difference based RECA-S, approximations and related measures

Difference metric based measures			
Algorithm	Measure	Threshold	Condition
CC-DRECA	$m_{cr}(x_i, C_m)$	C	$\lvert d_{cr}(x_i, C_m) - d_{cr}(x_i, C_l)\rvert \leq \epsilon_{cr}$
CF-DRECA	$m_{cr}(x_i, C_m)$	F	$\lvert d_{fz}(x_i, C_m) - d_{fz}(x_i, C_l)\rvert \leq \epsilon_{fz}$
CP-DRECA	$m_{cr}(x_i, C_m)$	P	$\lvert d_{pr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{pr}$
C-FP-DRECA	$m_{cr}(x_i, C_m)$	FP	$\lvert d_{fp}(x_i, C_m) - d_{fp}(x_i, C_l)\rvert \leq \epsilon_{fp}$
FC-DRECA	$m_{fz}(x_i, C_m)$	C	$\lvert d_{cr}(x_i, C_m) - d_{cr}(x_i, C_l)\rvert \leq \epsilon_{cr}$
FF-DRECA	$m_{fz}(x_i, C_m)$	F	$\lvert d_{fz}(x_i, C_m) - d_{fz}(x_i, C_l)\rvert \leq \epsilon_{fz}$
FP-DRECA	$m_{fz}(x_i, C_m)$	P	$\lvert d_{pr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{pr}$
F-FP-DRECA	$m_{fz}(x_i, C_m)$	FP	$\lvert d_{fp}(x_i, C_m) - d_{fp}(x_i, C_l)\rvert \leq \epsilon_{fp}$
PC-DRECA	$m_{pr}(x_i, C_m)$	C	$\lvert d_{cr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{cr}$
PF-DRECA	$m_{pr}(x_i, C_m)$	F	$\lvert d_{fz}(x_i, C_m) - d_{fz}(x_i, C_l)\rvert \leq \epsilon_{fz}$
PP-DRECA	$m_{pr}(x_i, C_m)$	P	$\lvert d_{pr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{pr}$
P-FP-DRECA	$m_{pr}(x_i, C_m)$	FP	$\lvert d_{fp}(x_i, C_m) - d_{fp}(x_i, C_l)\rvert \leq \epsilon_{fp}$
FP-C-DRECA	$m_{fp}(x_i, C_m)$	C	$\lvert d_{cr}(x_i, C_m) - d_{cr}(x_i, C_l)\rvert \leq \epsilon_{cr}$
FP-F-DRECA	$m_{fp}(x_i, C_m)$	F	$\lvert d_{fz}(x_i, C_m) - d_{fz}(x_i, C_l)\rvert \leq \epsilon_{fz}$
FP-P-DRECA	$m_{fp}(x_i, C_m)$	P	$\lvert d_{pr}(x_i, C_m) - d_{pr}(x_i, C_l)\rvert \leq \epsilon_{pr}$
FP–FP-DRECA	$m_{fp}(x_i, C_m)$	FP	$\lvert d_{fp}(x_i, C_m) - d_{fp}(x_i, C_l)\rvert \leq \epsilon_{fp}$

Table 2. Distance threshold based RECA-S, approximations and related measures

Threshold metric based measures			
Algorithm	Measure	Threshold	Condition
CC-TRECA	$m_{cr}(x_i, C_m)$	C	$d_{cr}(x_i, C_m) \leq \epsilon_{cr}$
CF-TRECA	$m_{cr}(x_i, C_m)$	F	$d_{fz}(x_i, C_m) \geq \epsilon_{fz}$
CP-TRECA	$m_{cr}(x_i, C_m)$	P	$d_{pr}(x_i, C_m) \geq \epsilon_{pr}$
C-FP-TRECA	$m_{cr}(x_i, C_m)$	FP	$d_{fp}(x_i, C_m) \geq \epsilon_{fp}$
FC-TRECA	$m_{fz}(x_i, C_m)$	C	$d_{cr}(x_i, C_m) \leq \epsilon_{cr}$
FF-TRECA	$m_{fz}(x_i, C_m)$	F	$d_{fz}(x_i, C_m) \geq \epsilon_{fz}$
FP-TRECA	$m_{fz}(x_i, C_m)$	P	$d_{pr}(x_i, C_m) \geq \epsilon_{pr}$
F-FP-TRECA	$m_{fz}(x_i, C_m)$	FP	$d_{fp}(x_i, C_m) \geq \epsilon_{fp}$
PC-TRECA	$m_{pr}(x_i, C_m)$	C	$d_{cr}(x_i, C_m) \leq \epsilon_{cr}$
PF-TRECA	$m_{pr}(x_i, C_m)$	F	$d_{fz}(x_i, C_m) \geq \epsilon_{fz}$
PP-TRECA	$m_{pr}(x_i, C_m)$	P	$d_{pr}(x_i, C_m) \geq \epsilon_{pr}$
P-FP-TRECA	$m_{pr}(x_i, C_m)$	FP	$d_{fp}(x_i, C_m) \geq \epsilon_{fp}$
FP-C-TRECA	$m_{fp}(x_i, C_m)$	C	$d_{cr}(x_i, C_m) \leq \epsilon_{cr}$
FP-F-TRECA	$m_{fp}(x_i, C_m)$	F	$d_{fz}(x_i, C_m) \geq \epsilon_{fz}$
FP-P-TRECA	$m_{fp}(x_i, C_m)$	P	$d_{pr}(x_i, C_m) \geq \epsilon_{pr}$
FP-FP-TRECA	$m_{fp}(x_i, C_m)$	FP	$d_{fp}(x_i, C_m) \geq \epsilon_{fp}$

Table 3. The measures for RECA-S, for example for the lower and upper approximations

Approximation	Distance	Value
Cr	$m_{cr}(x_i, C_m)$	1
Fz	$m_{fz}(x_i, C_m)$	μ_{C_m}
Pr	$m_{pr}(x_i, C_m)$	$\|d_{pr}(x_i, C_m)$
FP	$m_{fp}(x_i, C_m)$	$\|d_{fpr}(x_i, C_m)$

the lower and upper approximations have been calculated, the data objects it is possible to assign data objects to the clusters in the procedure of cluster centers recalculation. The measures for the lower and upper approximations have been presented in Table 3.

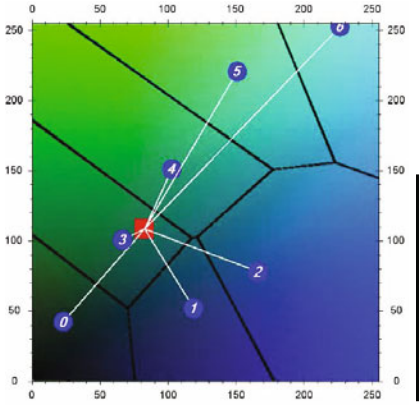
The example of the RECA-S set, RECA-CS cluster set and RECA-IA influence areas has been given in Fig. 1 (a) that are described by the cluster centers RECA-CC given in Fig. 1 (b).

2.3 The Distance Measures in Rough Extended Framework

Crisp RECA Measures. In crisp setting, *RECA* measures are calculated on the base of the crisp metric. Standard crisp distance most often applied in many working software data analysis systems depends upon Euclidean distance ($p = 2$) or arbitrary Minkowsky distance ($p > 0$), calculated as follows

$$d_{cr}(x_i, C_m) = \left(\Sigma_{j=1}^{d}(x_{ij} - C_{mj})^p\right)^{\frac{1}{p}} \tag{2}$$

Fuzzy *RECA* Measures. Fuzzy membership value $\mu_{C_l}(x_i) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as



a

ID	R	G	B
C0	0	41,5	23,1
C1	0	51,2	119,3
C2	0	77,4	166,8
C3	0	100,2	66,9
C4	0	150,9	103,2
C5	0	220,7	151,4
C6	0	252,9	227,2

b

Fig. 1. The color 0GB space with selected influence areas (a) of the cluster set RECA-S with the cluster centers RECA-CC given in (b)

$$d_{fz}(x_i, C_m) = \mu_{C_l}(x_i) = \frac{d(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d(x_i, C_j)^{-2/(\mu-1)}} \tag{3}$$

where a real number $\mu > 1$ represents fuzzifier value and $d(x_i, C_l)$ denotes distance between data object x_i and cluster (center) C_l.

Probabilistic *RECA* Measures. Probability distributions in *RECA* measures are required during measure calculations of probabilistic distance between data objects and cluster centers. Gauss distribution has been selected as probabilistic distance metric for data point $x_i \in U$ to cluster center C_m calculated as follows

$$d_{pr}(x_i, C_m) = (2\pi)^{-d/2} |\Sigma_m|^{-1/2} exp\left(-\frac{1}{2}(x_i - \mu_m)^T \Sigma_m^{-1}(x_i - \mu_m)\right) \tag{4}$$

where $|\Sigma_m|$ is the determinant of the covariance matrix Σ_m and the inverse covariance matrix for the C_m cluster is denoted as Σ_m^{-1}. Data dimensionality is denoted as d. In this way, for standard color RGB images $d = 3$, for gray scale images $d = 1$. Mean value for Gauss distribution of the cluster C_m has been denoted as μ_m.

Fuzzified Probabilistic RECA Measures. In fuzzified probabilistic *RECA* measures, the probabilistic distances to all clusters are fuzzified by means of the following formulae applied to the $d_1, \ldots, d_n$ distances. Fuzzified membership value of probabilistic distance $d_{fp}(x_i, C_l) \in [0, 1]$ for the data point $x_i \in U$ in cluster C_l is given as

$$d_{fp}(x_i, C_l) = \frac{d_{pr}(x_i, C_l)^{-2/(\mu-1)}}{\sum_{j=1}^{k} d_{pr}(x_i, C_j)^{-2/(\mu-1)}} \tag{5}$$

3 General Uniform *RECA* Transformations

In the present publication, a new generalized crisp RECA transform operator has been introduced in the domain of *RECA* measures denoted as $TR(RECA)$. The general *RECA* uniform transformation requires input image data *RECA*-I, arbitrary *RECA*-S set and *RECA* transform parameters - *RECA*-$P_{cr,fz,pr,fp}$. In this general case, cluster recalculation *RECA*-CR formulae is given as

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_r(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} m_r(x_p, C_k)} \tag{6}$$

In this context, *RECA* parameter set most often is denoted as

$$(C, F, P, FP) - (C, F, P, FP) - (T, D) - RECA \tag{7}$$

where Cr-C (crisp), Fz-F (fuzzy), P-Pr (probabilistic), FzPr-FP (fuzzified probabilistic) are used interchangeably. The uniform RECA-TR transformation is performed as the procedure given in Algorithm 1.

Algorithm 1. RECA-TR transformation

1. Take parameters for RECA-TR(P, C, I, S)
   (a) Calculate the following sets: RECA-CS, RECA-S, RECA-CA
   (b) For each data object x_i, calculate the membership value m_r in the following form

$$m_r = \{m_{cr}, m_{fz}, m_{pr}, m_{fp}, \}$$

   (c) Recalculate cluster centers

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_r(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} x_p * m_r(x_p, C_k)} \tag{8}$$

2. Perform data object re-assignment to clusters
3. Calculate the *RECA-TR*(P, C, I, S)
4. Repeat calculation of the *RECA-TR*(P, C, I, S) for the required number of iterations.

The term uniform in the context of RECA transformations means that the source transformed RECA-S set has the same type as resultant RECA-S. In this way, the four uniform transformations are possible

1. *RECA*(CrCr-TD)(CrCr-TD)
2. *RECA*(FzFz-TD)(FzFz-TD)
3. *RECA*(PrPr-TD)(PrPr-TD)
4. *RECA*(FzFz-TD)(FzFz-TD)

The detailed description of the all above-mentioned types of uniform *RECA* transformations has been given in the next Section.

Table 4. The parameters RECA-P

RECA-S Type	RECA-S Parameters	Name
P_{cr}	ϵ_{cr}	crisp threshold
P_{fz}	ϵ_{fz}	fuzzy threshold
	μ	fuzzifier
P_{pr}	ϵ_{pr}	probabilistic threshold
	$dist$	probabilistic distribution
P_{fp}	ϵ_{fp}	probabilistic threshold
	$dist$	probabilistic distribution
	μ	fuzzifier

4 Uniform *RECA* Transformations

4.1 $RECA_{cr}$ - Uniform RECA Crisp Transformation

In the present publication, a new generalized crisp *RECA* transform operator has been introduced in the domain of *RECA* measures denoted as

$$RECA(CrCr - TD)(CrCr - TD)$$

according to the notation in the Equ. 7. The first element (CrCr-TD) describes the crisp RECA-S set before the transformation. It means the Cr-TD describes the crisp threshold type in determining approximations and the first Cr describes the crisp approximation measure as described in Table 3. The second element (CrCr-TD) has the same description but denotes the resultant RECA-S set. The crisp *RECA* uniform transformation requires input image data *RECA*-I, crisp *RECA*-S set and crisp *RECA* transform parameters $RECA - P_{cr}$. In this case, the measure $m_r = m_{cr}$, calculated as in Equ. 2, *RECA*-CR formulae is given as

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_r(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} m_r(x_p, C_k)} \tag{9}$$

The parameters crisp $RECA - P_{cr} = \{\epsilon_{cr}\}$ as given in Table 4.

4.2 $RECA_{fz}$ - Uniform *RECA* Fuzzy Transformation

In the present publication, a new generalized fuzzy *RECA* transform operator has been introduced in the domain of *RECA* measures denoted as

$$RECA(FzFz - TD)(FzFz - TD)$$

according to the notation in the Equ. 7. The first element (FzFz-TD) describes the crisp RECA-S set before the transformation. It means the Fz-TD describes the fuzzy threshold type in determining approximations and the first Fz describes the fuzzy approximation measure as described in Table 3. The second element (FzFz-TD) has the same description but denotes the resultant RECA-S set.The fuzzy *RECA*uniform transformation requires input image data *RECA*-I, fuzzy

RECA-S set and fuzzy *RECA* transform parameters $RECA - P_{fz}$. In this case, the measure $m_r = m_{fz}$, calculated as in Equ. 3, *RECA*-CR formulae is given as

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_{fz}(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} m_{fz}(x_p, C_k)} \tag{10}$$

The parameters fuzzy $RECA - P_{fz} = \{\epsilon_{fz}, \mu\}$ are given in Table 4.

4.3 *RECA*$_{pr}$ - Uniform *RECA* Probabilistic Transformation

In the present publication, a new probabilistic operator has been introduced in the domain of *RECA* measures denoted as

$$RECA(PrPr - TD)(PrPr - TD)$$

according to the notation in the Equ. 7. The first element (PrPr-TD) describes the crisp RECA-S set before the transformation. It means the Pr-TD describes the probabilistic threshold type in determining approximations and the first Pr describes the probabilistic approximation measure as described in Table 3. The second element (PrPr-TD) has the same description but denotes the resultant RECA-S set. The probabilistic *RECA* uniform transformation requires input image data *RECA*-I, probabilistic *RECA*-S set and probabilistic *RECA* transform parameters $RECA - P_{pr}$. In this case, the measure $m_r = m_{pr}$, calculated as in Equ. 4, *RECA*-CR formulae is given as

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_{pr}(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} m_{pr}(x_p, C_k)} \tag{11}$$

The parameters probabilistic $RECA - P_{pr} = \{\epsilon_{pr}, dist\}$ are given in Table 4.

4.4 *RECA*$_{fp}$ - Uniform *RECA* Fuzzified Probabilistic Transformation

In the present publication, a new fuzzified probabilistic operator has been introduced in the domain of *RECA* measures denoted as

$$RECA(FPFP - TD)(FPFP - TD)$$

according to the notation in the Equ. 7. The first element (FPFP-TD) describes the crisp RECA-S set before the transformation. It means the FP-TD describes the probabilistic threshold type in determining approximations and the first FP describes the fuzzified probabilistic approximation measure as described in Table 3. The second element (FPFP-TD) has the same description but denotes the resultant RECA-S set. The fuzzified probabilistic *RECA* uniform transformation requires input image data *RECA*-I, probabilistic *RECA-S* set and fuzzified probabilistic *RECA* transform parameters *RECA-P*$_{fp}$. In this case, the measure $m_r = m_{fp}$, calculated as in Equ. 5, *RECA-CR* formulae is given as

$$C_m = \Sigma_{i=0}^{n} \frac{x_i * m_{fp}(x_i, C_m)}{\Sigma_{j=0}^{k} \Sigma_{p=0}^{n} m_{fp}(x_p, C_k)} \tag{12}$$

The parameters fuzzified probabilistic *RECA*-P_{fp}= $\{\epsilon_{fp}, dist, \mu\}$ are given in Table 4.

4.5 Uniform RECA Transformation Path

The concept of different properties of uniform *RECA* transformations has been the starting point in definition of the concept of *RECA*-paths. The RECA-paths is a new notion of transformation path geometrical properties and the rough properties such as entropies of the resultant RECA-S sets.

5 Conclusions and Future Research

In the study, the definition, detailed analysis and presentation material of the uniform *RECA* transformations have been presented. The *RECA* transformations are the part of the Rough Extended Clustering Framework designed into the unification of rough set theory in the clustering setting. The introduced solution seems to incorporate the main three k-means based clustering algorithms: standard k-means, fuzzy k-means and *EM* k-means clustering into one clustering framework. These three basic clustering algorithms seem to be a special case of Uniform *RECA* transformations.

The combination of the crisp, fuzzy, probabilistic and fuzzified probabilistic rough measures together with application of different notions based upon rough sets theory created robust theoretical framework in design, implementation and application of algorithmic procedures capable of high quality data segmentation. The application of rough approximations based upon rough transformations seems to be novel emerging approach in data analysis.

Acknowledgments

The research is supported by the Rector's grant of Bialystok University of Technology.

References

1. Malyszko, D., Stepaniuk, J.: Granular Multilevel Rough Entropy Thresholding in 2D Domain. In: 16th International Conference Intelligent Information Systems, IIS 2008, Zakopane, Poland, June 16-18, pp. 151–160 (2008)
2. Małyszko, D., Stepaniuk, J.: Standard and Fuzzy Rough Entropy Clustering Algorithms in Image Segmentation. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 409–418. Springer, Heidelberg (2008)

3. Malyszko, D., Stepaniuk, J.: Adaptive multilevel rough entropy evolutionary thresholding. Information Sciences 180(7), 1138–1158 (2010)
4. Malyszko, D., Stepaniuk, J.: Adaptive Rough Entropy Clustering Algorithms in Image Segmentation. Fundamenta Informaticae 98(2-3), 199–231 (2010)
5. Pal, S.K., Shankar, B.U., Mitra, P.: Granular computing, rough entropy and object extraction. Pattern Recognition Letters 26(16), 2509–2517 (2005)
6. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007)
7. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley & Sons, New York (2008)
8. Skowron, A., Stepaniuk, J.: Tolerance Approximation Spaces. Fundamenta Informaticae 27(2-3), 245–253 (1996)
9. Stepaniuk, J.: Rough–Granular Computing in Knowledge Discovery and Data Mining. Springer, Heidelberg (2008)
10. Slezak, D., Ziarko, W.: The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning 40, 81–91 (2005)

Another Variant of Robust Fuzzy PCA with Initial Membership Estimation

Gyeongyong Heo[1], Seong Hoon Kim[2], Young Woon Woo[3], and Imgeun Lee[4]

[1] Visual Media Center, Dong-Eui University, Korea
[2] Dept. of Software Engineering, Kyungpook National University, Korea
[3] Dept. of Multimedia Engineering, Dong-Eui University, Korea
[4] Dept. of Visual Information Engineering, Dong-Eui University, Korea

Abstract. Principal component analysis (PCA) is a well-known method for dimensionality reduction and feature extraction. PCA has been applied in many areas successfully, however, one of its problems is noise sensitivity due to the use of sum-square-error. Several variants of PCA have been proposed to resolve the problem and, among the variants, improved robust fuzzy PCA (RF-PCA2) demonstrated promising results. RF-PCA2, however, still can be affected by noise due to equal initial membership values for all data points. The fact that RF-PCA2 is still based on sum-square-error is another reason for noise sensitivity.

In this paper, a variant of RF-PCA2 called RF-PCA3 is proposed. The proposed algorithm modifies the objective function of RF-PCA2 to allow some increase of sum-square-error and calculates initial membership values using data distribution. RF-PCA3 outperforms RF-PCA2, which is supported by experimental results.

Keywords: principal component analysis, noise sensitivity, membership initialization, nearest neighbor, KD-tree.

1 Introduction

Principal component analysis (PCA) is a well-known and widely used method, which is optimal in the sense that it is an orthogonal transformation minimizing the sum of squared errors or reconstruction errors [1]. Although PCA has been used successfully in many applications, it has some problems and noise sensitivity is one of them. There have been several approaches to resolve noise sensitivity and they can be divided roughly into two groups: subset-based methods and fuzzy methods. Subset-based methods utilize one or more subsets of data to robustly estimate principal components (PCs) [2] [3] [4]. Although they showed successful results, they tend to suffer the small sample size problem and instability in calculating PCs.

To find robust PCs, fuzzy variants of PCA adopt fuzzy memberships to reduce the effect of outliers. Most fuzzy methods except robust fuzzy PCA (RF-PCA) consist of two steps, (1) estimating memberships and (2) finding PCs by building a fuzzy covariance matrix, and they put a focus on the first step [5] [6]. Fuzzy methods estimate memberships by formulating objective functions and

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 120–129, 2011.

optimizing them, however, the basic limitation is that only the first PC is used to estimate memberships. Although the first PC retains the largest portion of data variance, it can be easily affected by noise. Another problem is that the quantity optimized, memberships from fuzzy clustering for example, does not have a direct relationship with PCA.

RF-PCA, also belongs to the second group, extended previous methods by using $k(k \geq 1)$ PCs simultaneously and minimizing the sum of reconstruction errors in the estimation of memberships [7]. PCA minimizes the sum of reconstruction errors, therefore, it is natural to minimize the sum of reconstruction errors in the estimation of memberships. By iteratively optimizing memberships and PCs, RF-PCA demonstrated better result than other methods. However, using two different objective functions for memberships and PCs results in the lack of convergence property. The difference between two objective functions also slows the convergence and deteriorates the solutions of RF-PCA.

Improved robust fuzzy PCA (RF-PCA2) is a variant of RF-PCA in which the two objective functions are integrated to form a common objective function for memberships and PCs [8]. RF-PCA2 converges faster than RF-PCA and the solutions are closer to the desired ones than those of RF-PCA. However, RF-PCA2 still can be affected by outliers. The noise sensitivity comes from several reasons and assigning equal initial memberships for all data points is one of them. Another reason is that RF-PCA2 is still based on sum-square-error.

In this paper, a variant of RF-PCA2 called RF-PCA3 is introduced, which initializes memberships using Gaussian distribution and allows some increase of sum-square-error by introducing a new term to the objective function RF-PCA2. There are numerous available methods to estimate initial memberships [9], however, a simple method which assigns small membership values to outliers is used in this paper for a computational reason. Although small membership values may be assigned to some typical points, it can be corrected during the iterative optimization. By changing the initial memberships from equal values to data-dependent ones, RF-PCA3 has more chance to converge on a local optimum better than others.

As PCA comes from the minimization of sum-square-error, the dependence on sum-square-error cannot be overcome completely. By adding a term corresponding to the objective of PCA, however, the dependence can be relaxed. As the new term allows the objective function of RF-PCA3 can accommodate some increase of sum-square-error, RF-PCA3 can effectively reduce noise sensitivity.

In the next section, RF-PCA2 is briefly reviewed and RF-PCA3 is formulated in Section 3. Experimental results are given in Section 4 followed by a discussion.

2 RF-PCA2

RF-PCA2 is an iterative algorithm which tries to find k orthonormal basis vectors corresponding to robust PCs. Let $X = \{x_1, x_2, \cdots, x_N\}$ be a sample set and N is the number of data points. The objective function of RF-PCA2 can be written as

$$\begin{aligned}\arg\min_{W} \quad J &= \sum_{i=1}^{N} u_i ||(x_i - \mu_R) - WW^T(x_i - \mu_R)||^2 \\ &\quad + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \\ &= \sum_{i=1}^{N} u_i e(x_i) + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \\ \text{s.t.} \quad & W^T W = I,\end{aligned} \tag{1}$$

where u_i is the membership of x_i, μ_R is a weighted mean, σ is a regularization constant, W is a matrix having k basis vectors as columns, I is an identity matrix, and $e(x_i)$ is a reconstruction error. The first term measures the sum of membership-weighted reconstruction errors and the second one is a regularization term to make estimated PCs noise-robust.

Alternating optimization can be used to find an optimal W, in which $U = [u_1, u_2, \cdots, u_N]^T$ and W are updated iteratively. By taking a partial derivative of Eq. (1) with respect to u_i, one can obtain the update equation of memberships:

$$u_i = \exp\left(-\frac{e(x_i)}{\sigma^2}\right). \tag{2}$$

Similarly, by taking a partial derivative with respect to μ_R, one can obtain the update equation of a weighted mean:

$$\mu_R = \frac{\sum_{i=1}^{N} u_i x_i}{\sum_{i=1}^{N} u_i}. \tag{3}$$

Basis matrix W minimizing Eq. (1) under the constraint $W^T W = I$ can be obtained by finding k eigenvectors having k largest eigenvalues of a weighted covariance matrix, which is defined as

1: Initialize a membership vector $U_0 = [u_1, \ldots, u_N]^T = [1, \ldots, 1]^T$ and a counter $t = 0$.
2: **repeat**
3: $t \leftarrow t + 1$.
4: Calculate a weighted mean vector μ_R and a weighted covariance matrix C_R (Eqs. (3) and (4)).
5: Build an orthonormal basis matrix W_t using the eigenvectors of C_R.
6: Calculate the memberships U_t using Eq. (2).
7: **until** $J^{t-1} - J^t < \varepsilon$ or $t > t_{max}$
8: **return**

Fig. 1. RF-PCA2 algorithm

$$C_R = \frac{1}{N}\sum_{i=1}^{N} u_i (x_i - \mu_R)(x_i - \mu_R)^T. \tag{4}$$

The detailed derivation can be found in [8]. The RF-PCA2 algorithm is summarized in Fig. 1, where ϵ is a predefined constant and t_{max} is the maximum number of iterations.

3 RF-PCA3

Although RF-PCA2 outperformed previous methods, it still has some problems. First of all, as RF-PCA2 is based on sum-square-error, the result of RF-PCA2 still can be affected by outliers. This can be relaxed by adding a term corresponding to the objective of PCA, which results in another variant of robust fuzzy PCA called RF-PCA3. The new objective function can be written as

$$\begin{aligned} \arg\min_{W} \quad J &= \sum_{i=1}^{N} u_i ||(x_i - \mu_R) - WW^T(x_i - \mu_R)||^2 \\ &\quad + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i) \\ &\quad - \alpha \sum_{i=1}^{N} ||(x_i - \mu_R) - WW^T(x_i - \mu_R)||^2, \\ &= \sum_{i=1}^{N} (u_i - \alpha) e(x_i) + \sigma^2 \sum_{i=1}^{N} (u_i \log u_i - u_i), \\ \text{s.t.} \quad W^T W &= I, \end{aligned} \tag{5}$$

where $\alpha (\alpha > 0)$ is a weighting constant. The third term in Eq. (5) allows data points to have large reconstruction errors, which helps RF-PCA3 to reduce noise sensitivity due to the use of sum-square-error. The constant α balances the first and third term of Eq. (5) and, in this paper, the value of it was set as $\alpha = 0.5$ experimentally.

The update equations of RF-PCA3 can be obtained using the derivation in the previous section. RF-PCA3 also uses Eq. (2) to update memberships, however, basis matrix W is built using different weighted mean and weighted covariance matrix, which can be written as

$$\mu_R = \frac{\sum_{i=1}^{N} (u_i - \alpha) x_i}{\sum_{i=1}^{N} (u_i - \alpha)}, \tag{6}$$

$$C_R = \frac{1}{N}\sum_{i=1}^{N} (u_i - \alpha)(x_i - \mu_R)(x_i - \mu_R)^T. \tag{7}$$

Table 1. Sum of neighbor distances

SND	N_N	A	
D_1	N	I	Euclidean distance
D_2	N	$\text{cov}(X)$	Mahalanobis distance
D_3	K	I	K nearest neighbor, Euclidean distance

Another reason for noise sensitivity in RF-PCA2 is Step 1, initialization step in Fig. 1. RF-PCA2 uses uniform initialization, whereas, in this paper, we propose to use another membership initialization method that considers data distribution. As PCA assumes data follow a Gaussian distribution, a typical point is the one that lies in a dense region. In other words, a point that has a small sum of distances to its neighbors can be considered as typical and should have a large membership value. Equation (8) represents the sum of distances of x_i to its neighbors, called *sum of neighbor distances* (SND):

$$d_i = \sum_{j=1}^{N_N} ||x_i - x_{(j)}||_A^2, \qquad 1 \le i \le N, \tag{8}$$

where $N_N (N_N \le N)$ is the number of neighbors, $x_{(j)}$ is the j^{th} nearest neighbor of x_i, and A is a covariance matrix. In this paper, three SNDs using different N_N and A are considered. When $N_N = N$, all the data points participate in the calculation of SND and only a proper subset of X is used when $N_N < N$. When $A = I$, Euclidean distance is used as a distance measure and Mahalanobis distance is used when A is a data covariance matrix of $\{x_{(j)} | j = 1, \cdots, N_N\}$. Table 1 summarizes the differences among SNDs used in this paper. In Table 1, D_3 is the only local SND considering K nearest neighbors and the other two are global SNDs with Euclidean and Mahalanobis distance, respectively. Although D_1 is a simple and well-motivated measure for calculating initial memberships, it requires $O(N^2)$ time complexity, which is infeasible for large data. Therefore, D_3 was also considered because finding nearest neighbors can be done efficiently with KD-tree [10] [11]. D_2 is considered because Mahalanobis distance as a generalized Euclidean distance is better in handling Gaussian distributions than Euclidean distance in spite of its computational burden. However, Mahalanobis distance and K nearest neighbors was not considered in this paper because calculating N local covariance matrices requires too much computation.

Using Eq. (8), the initial membership in Step 1 of Fig. 2 can be calculated as

$$u_i = \exp\left(-\frac{d_i'}{\sigma_d}\right), \tag{9}$$

where σ_d is standard deviation of d_i. The value d_i' in Eq. (9) is a mapped distance defined as

$$d_i' = \max(d_i - \bar{d}, 0), \tag{10}$$

where $\bar{d}$ is an average of d_i. The RF-PCA3 algorithm is summarized in Fig. 2, which is different from RF-PCA2 in Step 1 and Step 4. In Step 1, a different

1: Initialize a membership vector U_0 using Eq. (9) and a counter $t = 0$.
2: **repeat**
3: $t \leftarrow t + 1$.
4: Calculate a weighted mean vector μ_R and a weighted covariance matrix C_R (Eqs. (6) and (7)).
5: Build an orthonormal basis matrix W_t using the eigenvectors of C_R.
6: Calculate the memberships U_t using Eq. (2).
7: **until** $J^{t-1} - J^t < \varepsilon$ or $t > t_{max}$
8: **return**

Fig. 2. RF-PCA3 algorithm

initialization method is adopted and in Step 4, weighted mean and weighted covariance are different from those of RF-PCA2.

4 Experimental Results

To investigate the effectiveness of the proposed method, RF-PCA2 and RF-PCA3 were implemented and tested using Matlab. Figure 3 shows a data set and the first PCs found by PCA and RF-PCA3. The data consist of 110 randomly generated points, 100 points from a Gaussian distribution and 10 noise points from another Gaussian distribution. The centers of two Gaussian distributions were given as $[0, 0]^T$ and $[4, 3]^T$, respectively, and the covariance matrices were $\Sigma_{data} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}$ and $\Sigma_{noise} = \begin{bmatrix} 0.33 & 0 \\ 0 & 0.33 \end{bmatrix}$. This data set will be referred to as G_1. As is clear from Fig. 3, PCA found a skewed PC due to noise, but RF-PCA3 found an unskewed one because the noise points have small membership values and are negligible in the calculation of the first PC.

The first PC found by RF-PCA2 is not depicted in Fig. 3 but it was a little different from that of RF-PCA3. The first experiment was, therefore, to compare the quality of PCs. The performance of each algorithm can be estimated by the angle between a PC from noise-free data and the corresponding PC from noisy data. Let w_0 be the first PC found by PCA using a noise-free data set that is

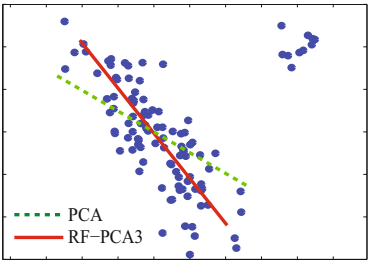
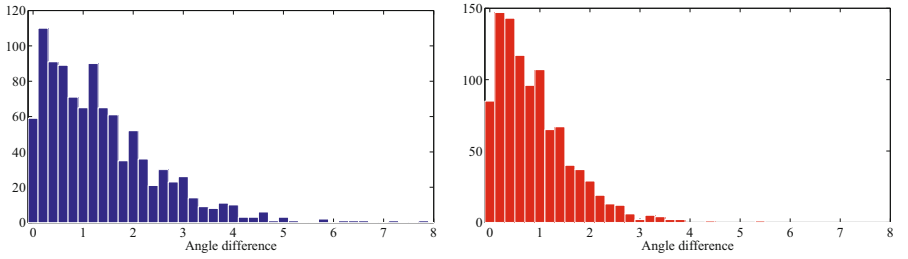


Fig. 3. Principal components found by PCA and RF-PCA3



(a) RF-PCA2 (b) RF-PCA3

Fig. 4. Angle histograms on G_1 using (a) RF-PCA2 and (b) RF-PCA3

equal to the data used in Fig. 3 without noise points. The vectors w_1 and w_2 are the first PCs found by RF-PCA2 and RF-PCA3 using a noisy data set, i.e., G_1. Figure 4 represents the histograms of angles between pairs of vectors – w_0 vs. w_1 and w_0 vs. w_2 – over 1,000 runs. The average angle between w_0 and w_1 is 1.357^o and that between w_0 and w_2 is 0.886^o, which means that PCs found by RF-PCA3 are more similar to the desired PCs than those of RF-PCA2 and that RF-PCA3 is more noise resistant than RF-PCA2. Even more, RF-PCA3 converges faster than RF-PCA2. The average number of iterations was 5.531 for RF-PCA3 and 6.623 for RF-PCA2.

In the previous experiments, RF-PCA2 also showed reasonable results. When the center of a noise distribution was moved from $[4, 3]^T$ to $[7, 7]^T$ (this data set will be referred to as G_2), however, RF-PCA2 fell into a local optimum as shown in Fig. 5. This is mainly from the uniform initialization in RF-PCA2. RF-PCA3, however, initializes memberships in a data-dependent way, which makes RF-PCA3 not to fall into a local optimum. Figure 6 summarizes experimental results on G_2. As is clear from Fig. 6, RF-PCA2 falls into a local optimum most of the time, whereas RF-PCA3 rarely falls into a local optimum.

RF-PCA2 and RF-PCA3 were also compared using another type of noise. Previous experiments used a Gaussian distribution to generate noise points, whereas, in this experiment, uniform distribution was used to sample noise, which will be referred to as G_3. Figure 7 shows the data and the first PCs found

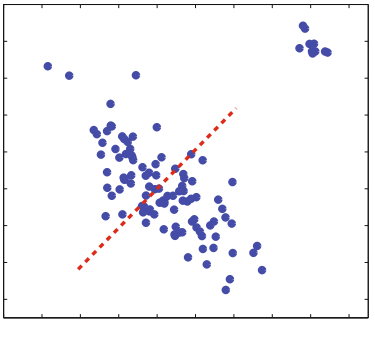
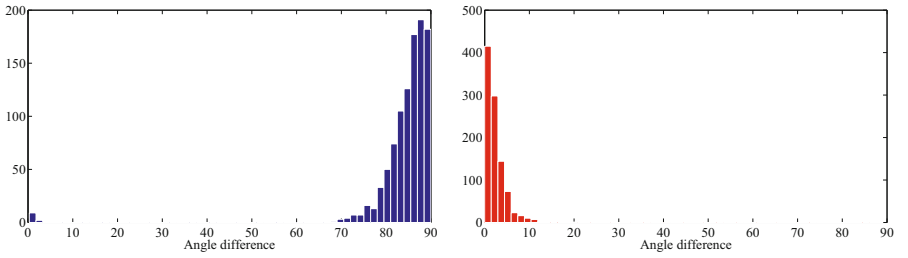


Fig. 5. Local optimum in RF-PCA2



(a) RF-PCA2 (b) RF-PCA3

Fig. 6. Angle histograms on G_2 using (a) RF-PCA2 and (b) RF-PCA3

by RF-PCA2 and RF-PCA3. Noise ratio, the number of noise points divided by the number of data points, represents the degree of noise and the noise ratio in Fig. 7 is 0.7. Figure 8 summarizes the experimental results on G_3 with noise ratio ranging from 0.1 to 1.0. As is clear from Figs. 7 and 8, RF-PCA3 is better than RF-PCA2 under uniform noise condition and the difference increases as

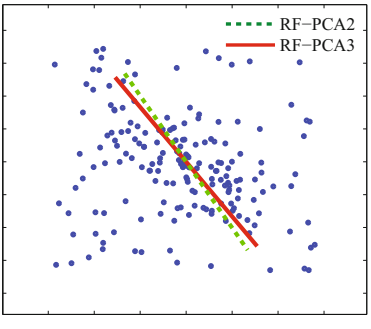


Fig. 7. Principal components found by RF-PCA2 and RF-PCA3

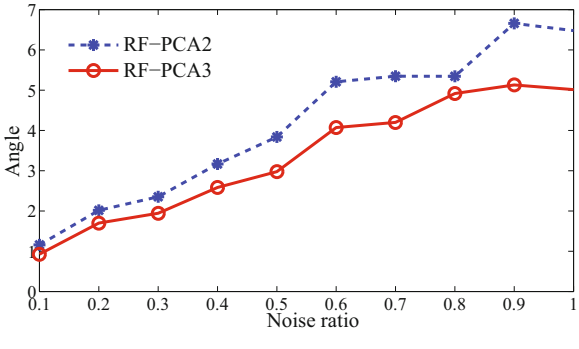


Fig. 8. Average angles with respect to noise ratio

noise ratio increases. This experiments also used the procedure described above except that the number of runs is 300 for each noise level. To calculate initial membership values D_1 was used up to here.

The last experiment was designed to compare the SNDs described in Table 1. Three SNDs were tested using both of the noise types. In this experiment, noise ratio was set as 0.6, however, experimental results with different noise ratio were almost identical to those presented here. Table 2 summarizes the experimental results with different SNDs. From this experiment, it is hard to say which one is better between D_1 and D_2 because none of the two consistently outperformed the other. However, D_3 outperformed the others in accuracy and speed all the time.

One problem in using D_3 lies in deciding K, the number of neighbors. When K is too small, the calculated value can be easily affected by noise. Whereas we cannot benefit from the fast nearest neighbor finding when K is too large. In this paper, K was set as 20 experimentally.

Table 2. Average angles and iteration numbers with respect to SND

	Gaussian blob noise		Uniform noise	
	Angle	Iteration Number	Angle	Iteration Number
D_1	2.3361	5.531	2.2226	6.950
D_2	2.2533	6.263	2.2257	6.534
D_3	2.1548	5.214	2.1788	6.370

5 Conclusions

PCA is a simple but effective method for dimensionality reduction and feature extraction. However, noise sensitivity originating from the objective of PCA is one of the problems in PCA. RF-PCA2 was the most promising one that uses fuzzy memberships and iterative optimization to mitigate noise sensitivity. RF-PCA2, however, is still affected by noise, which comes from two facts: uniform initialization and use of sum-square-error. In this paper, we proposed a variant RF-PCA2 which uses data-dependent initialization and modifies the objective function to allow some increase of sum-square-error. Experimental results using some artificial data sets showed that RF-PCA3 is more noise robust than RF-PCA2 although real-world applications will provide more thorough validation. We also tested three different initialization methods and a local SND considering some nearest neighbors was the best in speed and accuracy. However, deciding the number of neighbors is an unanswered question which is under investigation. Another problem in RF-PCA3 is its computational complexity. As RF-PCA3 is an iterative algorithm, it requires more computation than other methods. Another membership calculation method may reduces the computational complexity by accelerating the convergence without sacrificing performance, which is left for further research.

Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (Ministry of Education, Science and Technology) [NRF-2010-355-D00052] and the 2010 Culture Technology Joint Research Center Project of the Korea Creative Content Agency.

References

1. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, Heidelberg (2002)
2. Rousseeuw, P.J.: Multivariate estimation with high breakdown point. Mathematical Statistics and Applications B, 283–297 (1985)
3. Lu, C.-D., Zhang, T.-Y., Du, X.-Z., Li, C.-P.: A robust kernel PCA algorithm. In: Proceedings of the 3^{rd} International Conference on Machine Learning and Cybernetics, pp. 3084–3087 (2004)
4. Lu, C., Zhang, T., Zhang, R., Zhang, C.: Adaptive robust kernel PCA algorithm. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. VI 621–624 (2003)
5. Yang, T.-N., Wang, S.-D.: Fuzzy auto-associative neural networks for principal component extraction of noisy data. IEEE Transaction on Neural Networks 11(3), 808–810 (2000)
6. Cundari, T.R., Sarbu, C., Pop, H.F.: Robust fuzzy principal component analysis (FPCA). A comparative study concerning interaction of carbon-hydrogen bonds with molybdenum-oxo bonds. Journal of Chemical Information and Computer Sciences 42(6), 1363–1369 (2002)
7. Heo, G., Gader, P., Frigui, H.: RKF-PCA: Robust kernel fuzzy PCA. Neural Networks 22(5-6), 642–650 (2009)
8. Heo, G., Kim, S.H., Woo, Y.W.: An improved robust fuzzy principal component analysis. The Journal of the Korean Institute of Maritime Information and Communication Sciences 14(5), 1093–1102 (2010)
9. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22, 85–126 (2004)
10. Panigrahy, R.: An improved algorithm finding nearest neighbor using kd-trees. In: Proceedings of the 8th Latin American Symposium on Theoretical Informatics, pp. 387–398 (2008)
11. Nitin Bhatia, V.: Survey of nearest neighbor techniques. International Journal of Computer Science and Information Security 8(2), 302–305 (2010)

Automatic Emotion Annotation of Movie Dialogue Using WordNet

Seung-Bo Park[1], Eunsoon Yoo[1], Hyunsik Kim[1], and Geun-Sik Jo[2]

[1] Dept. of Information Engineering Inha Univ. Yonghyun-dong Nam-gu Incheon, Korea
{molaal,eunsoony,wbstory}@eslab.inha.ac.kr
[2] School of Computer & Information Engineering Inha Univ. Yonghyun-dong Nam-gu Incheon, Korea
gsjo@inha.ac.kr

Abstract. With the increasing interest in multimedia annotation, emotion annotation is being recognized as an essential resource which can be applied for a variety of purposes, including video information retrieval and dialogue systems. Here we introduce an automatic emotion annotation schema for dialogues, which could be used for the retrieval of specific scenes in film. Distinguished from previous works, we apply a new approach using the hypernym/hyponym relations and synonyms of WordNet, which enables us to organize a novel set of emotional concepts and to automatically detect whether a specific emotional word is associated with a specified emotional concept through measuring the conceptual distance between them.

Keywords: emotion annotation; WordNet; conceptual distance; emotional concept; emotion state.

1 Introduction

Since enormous volumes of multimedia contents are produced and distributed due to web development and advances in multimedia technology, users require quick and easy access to desired information. Therefore, annotating multimedia contents, which include a variety of semantic information, is indispensable for improving video information retrieval performance. Multimedia contents contain diverse significant features, from objective features such as visual objects to subjective features such as emotions. Most studies have concentrated on the recognition of objective features such as face recognition or audio pitch. However, since annotating subjective information has recently become available for multimedia retrieval and multimedia abstraction, advanced emotion annotation approaches have been researched.

In this paper, we describe an automatic emotion annotation schema for use in video information retrieval and video abstraction. As mentioned earlier, annotation plays an important role in advanced information retrieval applications. For example, emotion annotation in dialogues could be used to query specific sections in films [1]. People could search for particularly frequent or impressive scenes whenever necessary. Scenes in a film can be annotated by different emotion concepts such as ‘happiness’,

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 130–139, 2010.

'sadness', and 'anger', as actors and actresses express various emotions throughout the story. For example, a scene where the two main characters have a dispute can be labeled as 'anger' based on the use of words such as 'angry' and 'hate'.

To date, there has been considerable research on the relation between speech and emotion. The main issues are concentrated on the labeling of emotions and their recognition or detection based on choosing from lists of emotion concepts proposed in different disciplines.

The problems exposed by related works can be summarized as follows:

- **No definitive taxonomy of emotion:** There is no definitive taxonomy of emotions that is widely recognized by researchers though numerous emotion categories have been suggested in diverse fields from a minimal set of emotions such as Positive/Negative [4] to the big six categories (anger, fear, happiness, sadness, surprise, and disgust) [2, 3]. Therefore selecting a specific emotional label is difficult task. Besides, employing specific emotional labels lead to unsatisfactory coverage, because a great number of dialogues could be missed.
- **Manual annotation:** Manual annotation has a limitation. If a group of people annotate manually, they will achieve little consensus in determining and naming emotions [2, 3, 4, 5].

In order to overcome these problems, our work deals with the challenges of two main tasks:

- **New emotional concepts based on WordNet:** We propose a suitable large scale emotion taxonomy which allows the majority of dialogues to be annotated instead of existing emotion labels. To achieve this, we extracted 43 emotional concepts from WordNet, which are hyponyms of three emotional concepts, 'feeling', 'emotion', and 'emotional state'. In order to group these 43 emotional concepts as similar concept, we clustered them to 30 emotional concepts.
- **Automatic emotional annotation schema:** Differently from other research, we introduce the automatic emotional annotation schema using semantic relations such as hypernym/hyponym and synonym that are represented in WordNet. In addition, the Automatic Emotion Annotator (AEA) using this schema is implemented. After parsing dialogues by the Stanford POS Tagger [8], we extract nouns, adjectives, verbs and adverbs from each dialogue. We automatically detect which of the 30 emotional concepts a specific extracted word belongs to, and describe each word by points measuring the conceptual distance between a specific word and the relevant emotional concepts.

The rest of this paper is organized as follows. Section 2 reviews related works in emotion annotation. Section 3 describes our automatic emotion annotation schema. In section 4, we provide examples from our current experimental investigations. Section 5 summarizes our conclusions and outlines our ongoing work.

2 Related Works

The approaches adopted for describing emotional states can be largely divided into two types, specific emotion concepts and abstract dimensions.

Devillers [4] and Ling Chen [5] deal with the former. Devillers' schema [4] is used for labeling emotions in dialogues between agents and customers using five emotions: anger, fear, satisfaction, excuse and neutral. The annotation is accomplished by three annotators, who manually label each sentence with one of the five emotions. Since the list of emotions that Devillers selects is adapted for call center services, it is not clear that it would be equally applicable to different domains. EmoPlayer developed by Ling Chen is a media player used for playing a video clips with affective annotations [5]. It shows the emotions expressed by characters in films along a video timeline through color bars using mapping between emotions and colors. The various emotional expressions can be left unlabeled, because Ling Chen employs a simple emotion concept such as happy, sad, fearful, angry and neutral for affection annotation. As is the case for the schema of Devillers, emotions are labeled manually by a group of people.

Craggs proposed an annotation schema based on dimensional scales instead of adopting one of the specific emotion labels suggested in different disciplines [2, 3]. His schema is to annotate the emotional states in dialogues between nurses and cancer patients in terms of a space with two dimensions, intensity and evaluation. The intensity dimension describes the degree of emotion displayed in utterances, and each utterance can have a value from 0 to 4. The evaluation dimension labels the polarity of emotion expressions in utterances using three values, positive, negative and neutral. Craggs' schema also relies on manual annotation. One problem exposed by this schema is that there is no consistency among annotators, because it does not present criteria for determining the intensity of emotions in utterances.

The previous approaches are faced with the following two main problems. First, taking the specific emotion concepts gleaned in psychology and linguistics does not guarantee satisfactory reliability and high rates of coverage, because the majority of emotional expressions could be omitted. Second, an error can occur when annotating manually, and it is hard to achieve a result that can agreed upon by a number of annotators. Thus, we try to propose an automatic emotion annotation schema using WordNet.

3 Automatic Emotion Annotation Schema Using WordNet

Human emotion can be conveyed by various media such as facial display, gesture, speech, and a large vocabulary such as 'cry', 'nervous', 'gloomy', etc. Nowadays, since linguistic expression is available for studying emotion, several emotion annotation schemas for dialogue have been suggested.

Our approach is aimed at differentiating among the variety of emotion annotation schemata. We have organized a new set of emotional concepts by taking advantage of WordNet instead of selecting existing emotion labels. We extracted three concepts, emotional state, emotion and feeling from WordNet and the 43 sub-emotional concepts included. Besides, we made use of the semantic relations of WordNet for the automatic emotion annotation schema.

This schema can automatically perform a mapping between emotional words expressed in dialogues and emotional concepts using semantic structures such as hypernym/hyponym and synonym represented in WordNet. Furthermore, it can

describe an emotional state of a dialogue by points measuring the conceptual distance between an emotional word and the emotional concept that it belongs to.

3.1 Automatic Emotion Annotation Architecture

For each dialogue, we measure the emotional states in dialogues and annotate the emotional state. Our proposed schema consists of three phases as shown in Fig. 1:

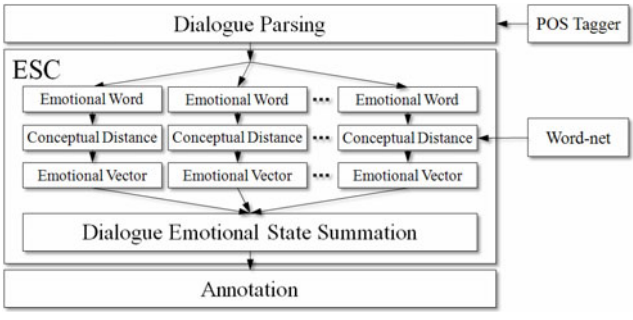


Fig. 1. Automatic Emotion Annotation Architecture

- **Dialogue Parsing:** A dialogue received from a subtitle is parsed by the POS tagger [8], and then, nouns, verbs, adjectives, and adverbs are extracted from a dialogue. Unemotional words or stop words, such as pronoun, 'be' verb, and article, are erased.
- **Emotional State Calculation (ESC):** An emotional state is measured by labeling automatically each emotional concept and calculating each emotional value for parsed words through the process, as shown on center of Fig. 1.
- **Annotation:** An emotional state is annotated in a dialogue.

3.2 WordNet-Based Emotional Concept

In this section, we present our WordNet-based emotional concept, which is built for gaining sufficiently high rates of coverage by labeling emotions in dialogues as efficiently as possible.

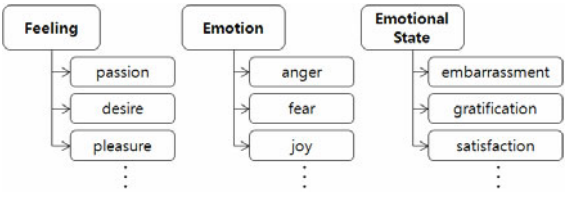


Fig. 2. Example of emotional concepts extracted from WordNet

Firstly, we selected three concepts, ‘emotional state’, ‘emotion’, and ‘feeling’ that are described in WordNet. These concepts have a hierarchical structure. The concept of ‘feeling’ is a super class of the concept of ‘emotion’, which is a super class of the concept of ‘emotional state’ in WordNet.

Secondly, we extracted 43 emotional concepts from WordNet, which are hyponyms of these three concepts. Fig. 2 shows examples of hyponyms taken from WordNet [7, 9]. We arranged the 43 emotional concepts by grouping them based on their definitions that are described in WordNet. For example, ‘pleasure’, ‘joy’, ‘enthusiasm’, and ‘satisfaction’ are grouped under the word ‘happiness’ because they contain the concept ‘happiness’ or ‘joy’ in the definition. Therefore, the word ‘happiness’ is hypernym of ‘pleasure’, ‘joy’, ‘enthusiasm’, and ‘satisfaction’.

Finally, we created a new set of 30 emotional concepts by grouping the 43 emotional concepts extracted from WordNet, as shown in Fig. 3.

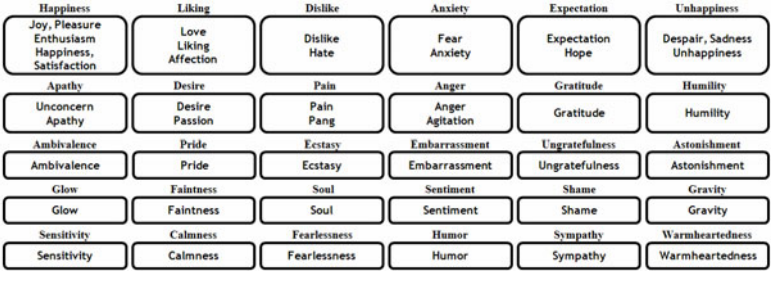


Fig. 3. Set of 30 emotional concepts to be grouped

3.3 Measuring Conceptual Distance Using WordNet

Notations used to calculate emotional state of a dialogue are stated in Table 1.

Table 1. Notations for the proposed method

Notations	Description
TECW	Top Emotional Concept Word
ESD_k	Emotional state of *k*th dialogue
$EV = <ev_1,ev_2,\ldots,ev_n>$	Emotional vector (Represents an emotional state)
ev_i	An emotional value for *i*th emotional concept
ev(word)	An emotional value of word (range of values: 1~8)
dist(word)	A conceptual distance of word. Its initial value is 0.

Two emotional vectors (EV_a, EV_b) are added for only the same emotional concept as given by equation 1.

$$EV_a+EV_b = <ev_{a1}+ev_{b1},ev_{a2}+ev_{b2},\ldots,ev_{an}+ev_{bn}> \quad (1)$$

Our automatic emotion annotation schema relies on the diverse semantic structures of WordNet. This is a large-scale lexical data base of English, which contains nouns,

verbs, adjectives and adverbs, and represents the meaning of each word by various semantic relations such as hypernym, hyponym, and synonym. This provides a suitable environment for annotating the emotions in dialogues in a fully automatic manner.

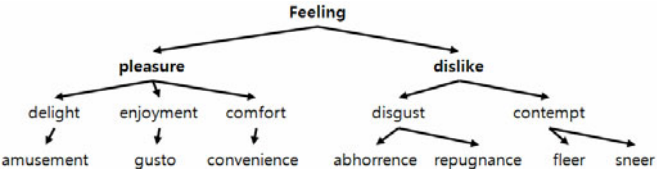


Fig. 4. Example of a hierarchy of the words 'pleasure' and 'dislike'

We take advantage of the hierarchy of hyponym/hypernym relations and synonyms for automatically labeling emotions in dialogue and measuring the conceptual distance between a specific emotional word and the emotional concept with which it is associated. Fig. 4 shows a hierarchy of the hypernym/hyponym relation of the words 'pleasure' and 'dislike'. For example, hypernym/hyponym relations and synonyms enable us to not only automatically find that an emotional word such as 'nice' can be labeled by the emotional concept 'pleasure' but also calculate the conceptual distance between the word 'nice' and the emotional concept 'pleasure'.

The conceptual distance is defined in [6] as "*the length of the shortest path that connects the concepts in a hierarchical semantic net*". In our research, it provides the basis for determining the closeness in meaning between a specific emotional concept and a specific emotional word. The closer a specific word is to the related emotional concept, the higher the level of semantic relatedness between them. On the contrary, the greater the distance of a specific word from the related emotional concept, the lower the level of semantic relatedness between them. As shown in Fig. 5, the further a specific word is from the relevant emotional concept, the more points are deducted. 2 points are deducted for a hyponym and 1 point is deducted for a synonym.

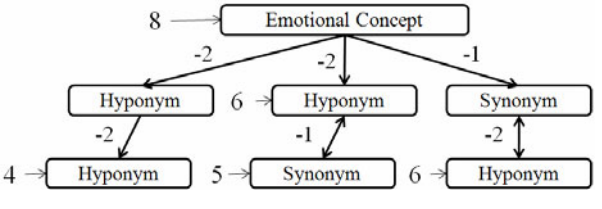


Fig. 5. Conceptual distance

Since the higher the value of the conceptual distance between a specific word and the related emotional concept, the higher the level of semantic closeness between them, we take the highest one from among the values obtained.

A Top Emotional Concept Word (TECW), such as happiness, liking, and dislike, has an emotional value of 8. The conceptual distance (dist) from the word to the TECW is calculated by equation 2 and 3. For a hypernym connection on the path

from TECW to a specific word, the distance of this word increases 2 points by equation 2.

$$dist_{k+1}(word) = dist_k(word) + 2 \quad (2)$$

For a synonym connection on the path, it increases 1 point by equation 3.

$$dist_{k+1}(word) = dist_k(word) + 1 \quad (3)$$

An emotional value (ev) of a word is obtained from equation 4 that subtracts the conceptual distance from the emotional value of the TECW.

$$ev(word) = ev(TECW) - dist(word) \quad (4)$$

Since the calculated emotional value of a word means the size of the *i*th emotional concept, an emotional word can be represented as a vector with the value at the *i*th element. Each element of this vector refers to an emotional concept, and the number of elements is 30, as shown in Figure 3. This vector is the emotional vector. Also, each emotional vector of a word is summed, and then, the emotional vector of a dialogue is calculated using equation 1. This emotional vector represents the emotional state of a dialogue.

The calculation of emotional state for a dialog is represented as Algorithm 1.

Algorithm 1. Calculation of emotional state for a dialog

```
Input : D                          //D: a dialog
Output : ESD                       //ESD: emotional state of a dialog D
ev = EmotionalVector(w)    // Function to calculate emotional vector of word (w)

function ESC(D)                //Emotional State Calculating Function
begin
  w[] = get words via parsing with POS-Tagger
  for i ← 1 to n
    ESD = ESD + EmotionalVector(w[i], 8) // 8 : initial value of conceptual distance of ith word
  return ESD
end

Input : w, score                   // w : Word, score : distance variable of word
Output : EV                        //EV : emotional vector of word (w)
Initial value of maxScore = 0
function EmotionalVector(w, score)
begin
  SynSet[] = get synonyms of w from WordNet
  for i ← 1 to n                   // n : the number of synonyms
    maxScore = Max(maxScore, EmotionalVector (SynSet[i], score-1))     // recursive call
    HypSet[] = get hypernyms of w from WordNet
    for j ← 1 to m                 // m : the number of hypernyms
      if HypSet[j] is TECW then
        if score > maxScore then
          maxScore = score         //update maxScore
      else
        maxScore = Max(maxScore, EmotionalVector (HypSet[j], score-2))   //recursive call
    set element of TECW at emotional vector (EV) as maxScore
  return EV
end
```

For example, let this process to be applied to "Honey! Don't worry. I will come back soon." This dialogue is sent to the function ESD to calculate emotional state of it as shown in Algorithm 1. At the Dialogue Parsing phase as shown in Fig. 1, the words 'honey (noun)' and 'worry (verb)' are extracted as **w**[] after the unemotional words are removed. At the next phase, the emotional vector of this dialogue is set by calculating the conceptual distance of each extracted word using the function EmotionalVector as shown in Algorithm 1. This function calls recursively to search a synonym or a hypernym on the path to TECW. This search is processed as follows.

- Honey >> lover-synonym(+1) >> love-derivation(+0): dist(honey) = 1 → ev(honey) = 8 – dist(honey) = 7
- worry >> anxiety-hypernym(+2): dist(worry) = 2 → ev(worry) = 8 – dist(worry) = 6

The word 'honey' is connected to 'love' as a synonym. Since emotional word 'honey' has only one synonym to the emotional concept of 'love', the emotional value of 'honey' is 7. Also, through the same process, the emotional concept of 'worry' is searched as 'anxiety' with a value of 6.

Therefore, an emotional vector (**EV**) for each word is represented using these concepts and values as follows:

- honey: <0,7,0,0,0,0,…,0>
- worry: <0,0,0,6,0,0,…,0>

The emotional state of the dialog "Honey! Don't worry. I will come back soon." (**ESD**) is calculated by the summation of two vectors as follow by equation 1.

- Honey! Don't worry. I will come back soon.: <0,7,0,6,0,0,…,0>

The calculated emotional state is annotated at this dialogue.

4 Evaluation and Discussion

In this section we present the experimental results of our automatic emotion annotation in dialogues of subtitles and evaluation. We selected 4 romantic comedies (M1 – M4) for evaluation, as shown in Table 2. For practical reasons, this is a sensible choice, since the appeal of these films is in the dramatic reality of the emotions expressed by the characters.

We implemented the AEA (Automatic Emotion Annotator) that detects and annotates the emotional states of dialogs with JAVA. We applied the Stanford POS-tagger version 3.0 for parsing dialogues to the AEA. Also, the AEA is implemented

Table 2. Information of evaluation data

Movie ID	Movie Name	Genre	# of Dialog
M1	He's just not that into you	Romance/Comedy	1098
M2	Notting Hill	Romance/Comedy	937
M3	Pretty Woman	Romance/Comedy	979
M4	You've got mail	Romance/Comedy	1147

by using a WordNet API (JAWS) for JAVA in order to calculate the conceptual distance in WordNet (version 2.1).

Ground truth: For each movie, we asked three specialists in film and literature to manually label emotional concepts and values for each dialog after watching movies. The emotional concepts that were labeled by all three persons were treated as the ground truth. And then, the average of the values labeled by the three persons was applied to the emotional value. However, if one person labeled a different emotional concept, this was removed.

4.1 Evaluation of Emotion Annotation

The following experiment investigates the effectiveness of WordNet-based emotion annotation. Table 3 summarizes the result of emotion annotation, showing how WordNet-based annotation improved the performance. We compared WordNet-based annotation with annotation based on the general big six categories (anger, fear, happiness, sadness, surprise, and disgust) for 4 movies (M1 – M4).

Table 3. Accuracy of total emotion annotation

Movie ID	WordNet-based (Success/TND*)	Big six (Success/TND*)
M1	69.70 %	47.47 %
M2	70.37 %	37.04 %
M3	78.00 %	28.00 %
M4	78.95 %	63.16 %
Average	74.25 %	43.92 %

TND*: The total number of dialogues.

The proposed WordNet-based method achieves a 30.34% average improvement for emotion annotation, compared to annotation based on the big six categories. While the big six method showed poor annotation performance, because it only used the six emotional categories, the WordNet-based method achieved excellent annotation performance, since various emotional concepts are detected.

The big six categories cannot be adapted to automatically annotate emotion in dialogues, since this method has a poor performance of less than 50%. In contrast, the WordNet-based method that achieves 74.25% average accuracy is efficient enough to automatically annotate emotion in dialogues.

Unlike the big six categories, we observed that numerous dialogues were annotated through the WordNet-based method. However, emotion annotation had a failure rate of approximately 26%, for the following reasons:

- **Polysemy:** When an emotional word has several senses, our schema might provide a wrong emotion annotation, because it chooses a sense that has the highest emotional value. For example, for the emotional word ‘cool’, our schema selected the sense ‘coolheaded’ instead of the sense ‘attractive’ because the former sense has higher emotional value than the latter one (e.g., good, cool, crappy, excuse).

- **No path to TECW:** In spite of the emotional word, there is no path to TECW in WordNet (e.g., babe, stupid, yep, fool).
- **Idiom with emotion** (e.g., 'of course', 'for god's sake', 'out of business')

We need additional researches in order to overcome these problems. To solve the polysemy problem, more research to select one sense adequate to the context among several senses is needed. For the no path and emotional idiom problem, a function to assign a synonym pair is also required.

5 Conclusion

In this paper we have described an Automatic Emotion Annotator (AEA) for annotating expressions of emotion in dialogues, and measuring the conceptual distance between a specific emotional word and a related emotional concept. We recognized that WordNet-based emotion concepts have the advantage of allowing numerous dialogues to be annotated rather than labeling emotion using the big six general categories through experimental comparison. Also, we have identified that the semantic relations of WordNet provide an effective environment for annotating the emotion in dialogues automatically.

For developing a general purpose schema for expressions of emotion in dialogue, the next step of our research will be to resolve previously mentioned problems such as processing of polysemy and collocations.

References

1. Chan, C.H., Jones, G.J.F.: Affect-Based Indexing and Retrieval of Films. In: 13th Annual ACM International Conference on Multimedia, pp. 427–430 (2005)
2. Craggs, R., Wood, M.M.: A Categorical Annotation Schema for Emotion in the Linguistic Content of Dialogue. In: The Affective Dialogue Systems, Tutorial and Research Workshop, pp. 89–100 (2004)
3. Craggs, R., Wood, M.M.: A two dimensional annotation scheme for emotion in dialogue. In: AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford University, pp. 44–49 (2004)
4. Devillers, L., Vasilescu, I., Lamel, L.: Annotation and detection of emotion in a task-oriented human-human dialog corpus. In: ISLE Workshop on Dialogue Tagging for Multi-Modal Human-Compute Interaction, pp. 15–17 (2002)
5. Chen, L., Chen, G.-C., Xu, C.-Z., March, J., Benford, S.: EmoPlayer: A media player for video clips with affective annotations. J. Interacting with Computers 20(1), 17–28 (2007)
6. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. J. IEEE Trans. on Systems, Man and Cybernetics 19(1), 17–30 (1989)
7. Richarson, R., Smeaton, A.F., Murphy, J.: Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words. In: AICS Conference (1994)
8. Stanford POS Tagger, `http://nlp.stanford.edu/software/tagger.shtml`
9. Miller, G.A.: WordNet: A Lexical Database for English. J. Communications of the ACM 38(11), 39–41 (1995)

Self–Organizing Map Representation for Clustering Wikipedia Search Results

Julian Szymański

Department of Computer Systems Architecture,
Gdańsk University of Technology, Poland
julian.szymanski@eti.pg.gda.pl

Abstract. The article presents an approach to automated organization of textual data. The experiments have been performed on selected sub-set of Wikipedia. The Vector Space Model representation based on terms has been used to build groups of similar articles extracted from Kohonen Self-Organizing Maps with DBSCAN clustering. To warrant efficiency of the data processing, we performed linear dimensionality reduction of raw data using Principal Component Analysis. We introduce hierarchical organization of the categorized articles changing the granularity of SOM network. The categorization method has been used in implementation of the system that clusters results of keyword-based search in Polish Wikipedia.

1 Introduction

The amount of information given in the form of documents written in a natural language requires researching methods for effective content retrieval. One way of improving retrieval efficiency is performing documents categorization which organizes documents and allows find relevant content easier.

In the article we present an approach to organization of a set of textual data through an unsupervised machine learning technique. We demonstrate how our method works on test dataset and describe the system that utilizes the method for categorization of the search results retrieved form Wikipedia.

Because of high dimensionality of the processed data (documents represented with terms as their features) we used a statistical method of **P**rincipal **C**omponent **A**nalysis [1]. The method identifies significant relations in the data and combines correlated features into one artificial characteristic which allows to reduce features space significantly. In the reduced feature space we construct Kohonen **S**elf-**O**rganising **M**ap [2] which allows 2D presentation of topological similarity relations between objects (here documents). Employing DBSCAN clustering we extract from the SOM groups of the most similar documents. Changing the SOM granularity we construct hierarchical categories that organize documents set.

2 Text Representation

The documents, to be effectively processed by machines, require to be converted from the form readable to humans into a form processable by machines. The main problem is

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 140–149, 2011.

a drawback between text representation used by humans, and that of the machines. Humans, while reading the text, understand its content, and thus he or she is able to know what it is all about. Despite some promising projects, understanding of a text by machines is still unsolved [3], [4]. Because machines don't understand the language they use features for document description which allowes extraction of important relations between processed data. In Artificial Intelligence it is called knowledge representation, and it aims at presenting some aspects of the world in a computable form [5].

In Information Retrieval [6] a typical approach for text representation [7] is usage of Vector Space Model [8] where documents are represented as points in feature space. As features typically words or links are used.

In the experiments presented here we use document content to represent it. This text representation employs words which the article contains. The features set has a size near to the number of all distinctive words which appear in the processed repository of the documents. To reduce size of the set we perform text preprocessing which contains the following procedures:

- stop words removal – all words which appear in so-called stop words list are removed from the features set. This allows us to exclude words which are not very informative in terms of machine processing, and which bring noise to the data.
- stemming – this preprocessing procedure allows to normalize words, through bringing different inflections of the word into its basic form. As a result, different forms of the word are treated as the same term.
- frequency filtering – we remove terms that were related to only one document.

The words preprocessed in this way are called terms. The value, or descriptiveness of a term for a given document may be estimated by the strength w of association between the term and the text. Typically for n-th term and k-th document w value is calculated as a product of two factors: term frequency tf and inverse document frequency idf, given by $w_{k,n} = tf_{k,n} \cdot idf_n$. The term frequency is computed as the number of its occurrences in the document and is divided by the total number of terms in the document. The frequency of a term in a text determines its importance for document content description. If a term appears in the document frequently, it is considered as more important. The inverse document frequency increases the weight of terms that occur in a small number of documents. The idf_n factor describes the importance of the term for distinguishing documents from each other and is defined as $idf_n = \log(k/k_{term(n)})$, where k is the total number of documents, and $k_{term(n)}$ denotes the number of documents that contain term n.

3 The Data

To perform experiments which would validate our approach to organize search results in repositories of documents we find Wikipedia very useful. This large source of human knowledge contains articles referenced one to another and provides also a system of categories. Despite the fact that the Wikipedia category system is not perfect, it can be used as a validation set for algorithms which perform articles organization in an automated way.

Table 1. The data used in the experiments

Category name	Symbol and color	Number of articles
Biology	magenta □	66
Chemistry	green +	229
Mathematic	blue ·	172
Theology	black ∗	135

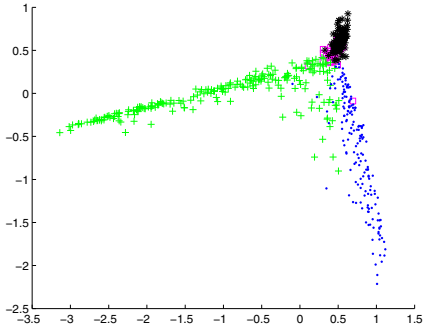


Fig. 1. View of the class distribution of the test dataset performed in 2D, created with two highest principal components

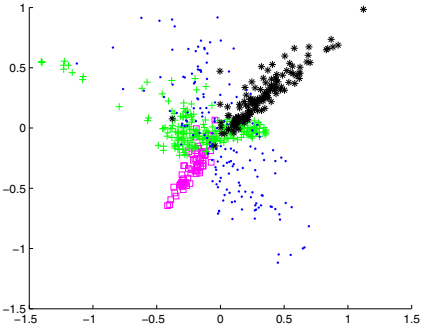


Fig. 2. View of the class distribution of the test dataset performed in 2D, where dimensions are created with third and fourth components

Wikipedia off-line data is publicly available for download[1]. The data contain SQL-dumps which provide structural information – linkages between articles and category assignments. There are also available XML dumps which offer textual content of all articles. Importing the data into local database and building the application allowed to extract selected information from XML files and turn it into computationable form. The application we have implemented allows us also to select articles for which the representation will be generated. It allows us to select only a subset of Wikipedia which warrants that the experiments run on single PC will be performed in reasonable time.

The experiments presented here we performed on test set of Wikipedia articles selected using categories. The articles in the set are relatively similar (they all belong to one super category). It enables to show usability of the presented method for introducing organization in documents set that contains elements that are conceptually similar. For test set we selected 602 articles which belong to 4 arbitrarily selected categories from one super category Science. In Table 1 we present categories used in the experiments, and the amount of the articles they contain. The initial features set consists of 37368 features that after preprocessing have been reduced to 12109.

[1] http://download.wikimedia.org

It is comfortable to have a rough view of the data. To see how it is distributed we provide visualization in 2D using principal components computed with PCA (described in section 4.1). What can be seen in Figure 1 the articles from category *biology* described with magenta □ are not separable from other classes. This task can be performed using other components what presents Figure 2 where third and fourth principal components have been used.

4 Documents Organization Method

Having documents represented with text representation, we are able to process them and perform experiments aiming at research methods for organizing them. Our approach we based on categorization and perform it in three steps:

1. dimension reduction,
2. mapping the the articles into Kohonen map,
3. exploiting proximities extracted from the map creation of articles clusters.

4.1 Dimension Reduction

The raw data we process are high dimensional (test set contains 12109 unique features). Some of the features which are used to describe processed objects are strongly correlated to one another. They can be replaced with artificial characteristic which is the combination of the original ones. One way of performing this a task is statistical method called **P**rincipal **C**omponent **A**nalysis [1]. The idea of the method is based on identification of principal components for the correlation matrix of the features. Selecting the most significant components is performed by computing eigenvectors of the correlation matrix and sorting them according to eigenvalues. A chosen number of eigenvectors which have the highest variance, can be used for representation of the original data. Multiplication of the truncated eigenvectors matrix by original data constructs lower dimensional space for representation of original objects. Selecting the number of eigenvectors used for reduction is crucial to obtain a good approximation of the original data. Very good approximation is to take eigenvectors that complete 99% of the data variance. In Figure 3 we present the % of variances for each of components we also provide information about the number of components whose cumulative sum completes 99% of the variance (136).

4.2 Self - Organizing Maps for Topological Representation of Articles Similarity

One of the methods for presenting significant relationships in the data is identification of similar groups of objects and, instead of the objects per se, presentation their representatives – prototypes. In our experiments we we use neural-inspired approach of the **S**elf-**O**rganizing **M**ap introduced by Kohonen [9]. The method is based on an artificial neural network which is trained in a competitive process, also called vector quantization [10]. The learning process uses the strategy **W**inner **T**akes **A**ll (in some algorithm versions **M**ost) which updates the weights of the neuron which is the most similar to the object used to activate the network. The neurons with strongest activations for the

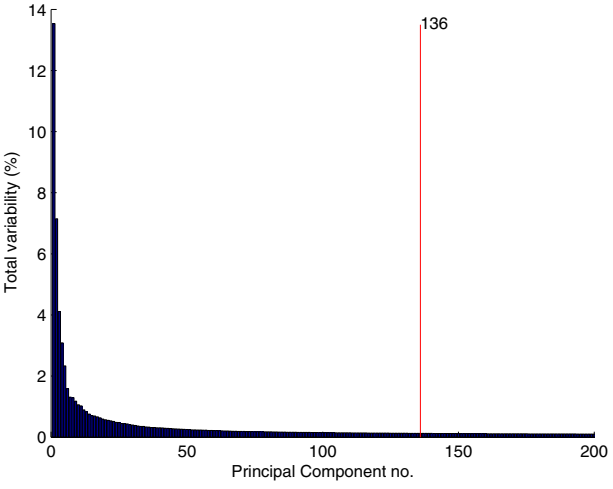


Fig. 3. The % of variability for succeeding principal components

objects that belong to the same class form prototypes [11] which can be used for representation of the particular set of objects.

The effect achieved after training the neural network, is functionally equal to nonlinear scaling from the n-dimensional objects to the smaller, here 2-dimensional space of their prototypes [12]. The advantage of the SOM method is its ability of graphical presentation. The results are visualized in 2D called maps where prototypes of the objects are presented. They keep topological distances according to the given object similarity measure which has been used during training of the neural network.

The SOM-based approach is known to be successfully applied in text processing [13] and it also found applications in web pages organization [14]. In this approach, the information retrieval process is accomplished with presentation of similarity of documents on the Kohonens map which aims at improving the searching process based on their proximity. The data presented in Table 1 and reduced with 136 highest principal components have been used to construct SOM. It presents topological relations between documents projected on 2D space where they are represented by their neural prototypes. SOM presented in (Figure 4) shows firing areas of the network while it was activated with articles that belong to different categories. We also provide joint SOM activation (figure 5) where areas of the network that overlaps have been marked with red.

4.3 Clustering the Data

The neurons of SOM may be interpreted as prototypes for articles. While the network is activated by articles that belong to different categories it responds with firing neurons from different areas of SOM. If articles belong to same category they activate close SOM areas. This fact allows to capture proximities of the articles on higher level of abstraction that arises form the fact the comparison is performed in low dimensional space (2 dimensions of SOM). The proximities can be computed calculating average

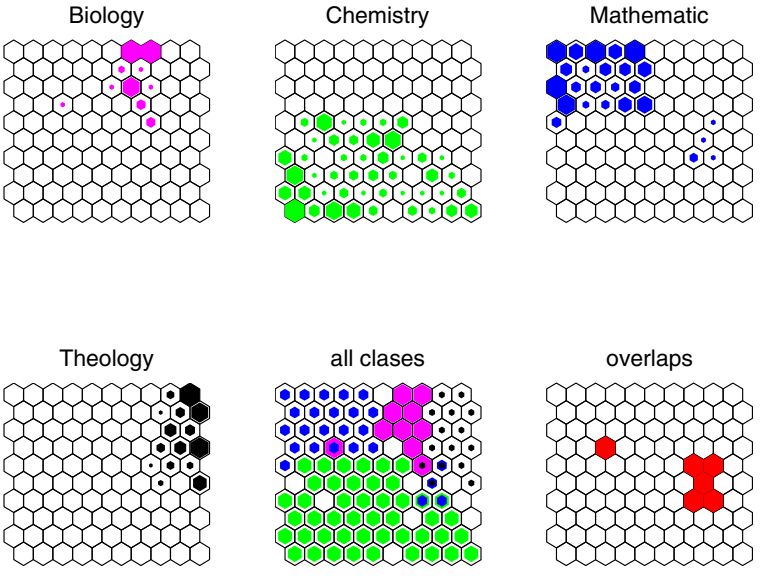


Fig. 4. Sample dataset presented on Self-Organized Map. Activations of different SOM areas for articles from different categories and their overlaps.

distances between neurons activated for each pair of the articles. It allows to construct articles similarity matrix where elements are calculated using formula 1.

$$sim(a_1, a_2) = \frac{1}{d(X,Y)} = \frac{1}{|X|}\sum_{i=1}^{|X|} d(x_i, Y) \tag{1}$$

where X and Y denote sets of neurons on SOM activated respectevely for article a_1 and a_2 and $d(x_i, Y)$ is calculated using formula 2.

$$d(x_i, Y) = \frac{1}{|Y|}\sum_{j=1}^{|Y|}\sqrt{(x_{i1} - y_{j1})(x_{i2} - y_{j2})} \tag{2}$$

The articles proximity matrix allows us to extract groups of the most similar articles. There are many methods and strategies to perform such a task [15]. We used here density-based approach that is known effective non parametric clustering technique suitable for textual data[16]. *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) [17] is a clustering algorithm based on densities of points in feature space. Its advantage is not very sensitive to noise and also it is able to find not only convex clusters, which is big limitation of typical clustering algorithms. The main idea of the algorithm is a concept of point neighborhood given by the radius ϵ (that is algorithm parameter) that must contain fixed, minimal number of other points (τ) belonging to the same cluster. The shape of neighborhood depends on proximity function. Eg.: for Manhattan distance it is rectangle. In our approach usage of formula 1 allows to build clusters of any shape. In DBSCAN there are three types of points: root (inside

cluster), border and outlayers. Changing the parameters ϵ and τ we can minimize number of outlayers and thus tune the algorithm. Border points are interpreted as articles that belong to more than one cluster and thus multi-categorisation is introduced, which is closer to the real-word categorization, performed by humans.

The clustering quality Q have been evaluated comparing cardinality of clusters C that has been computed and articles categories K created by humans. For each category K_j the quality Q_j is calculated using the formula 3.

$$Q_j = \frac{1}{|K|} \sum_{j=1}^{|K|} \frac{|a| \in C_{max}}{|a| \in K_j} \tag{3}$$

where $|a| \in C_{max}$ denotes number of articles that belong to cluster with highest cardinality and $|a| \in K_j$ denotes cardinality of K_j category.

For the categories from the test dataset (described in table 1) we obtain qualities shown in table 2.

Table 2. Clustering qualities for each of the category from sample dataset

Quality \ Category name	Biology	Chemistry	Mathematic	Theology
Q_j	0.71	0.82	0.84	0.66

4.4 Hierarchical Organization

Hierarchy is one of the most well-known ways for organization of large number of objects. It can also be built for a set of the documents using SOM [18], [19]. This task can be done changing the size of the SOM, which introduces different granularity of data organization. The organization is based on hierarchically layered prototypes that represent SOM neurons. The Figures 6 and 5 show bottom-up process of changing the size of the SOM which transforms articles to their more general prototypes. If we take prototypes that bind together sets of the articles, this process can be seen as generalizing the articles into more abstract categories. Overlaps between neuron activations induce the ability to introduce new relations between categories, as well as it show some other possible directions in which the categorization can be performed further.

5 Application and Future Directions

In the article we present the approach for documents categorization based on terms VSM used for representation of Wikipedia articles. We perform linear dimensionality reduction based on PCA. It allows to build Self-Organizing Map in effective way. Changing the SOM size we introduce hierarchical organization of the documents set. Using SOM representation for DBSCAN clustering we identify clusters of the most similar articles. We present how our method works for arbitrary selected categories of articles and how it allows to separate them.

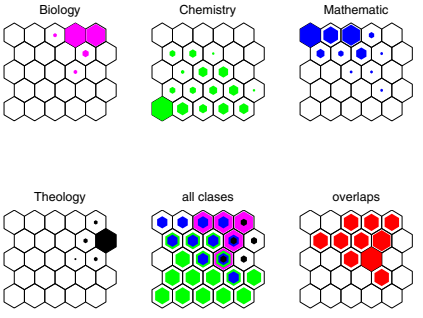


Fig. 5. SOM 5 *times* 5, for creating hierarchical categories

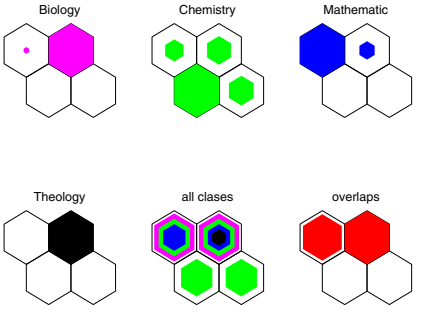


Fig. 6. SOM 2 *times* 2, for creating hierarchical categories

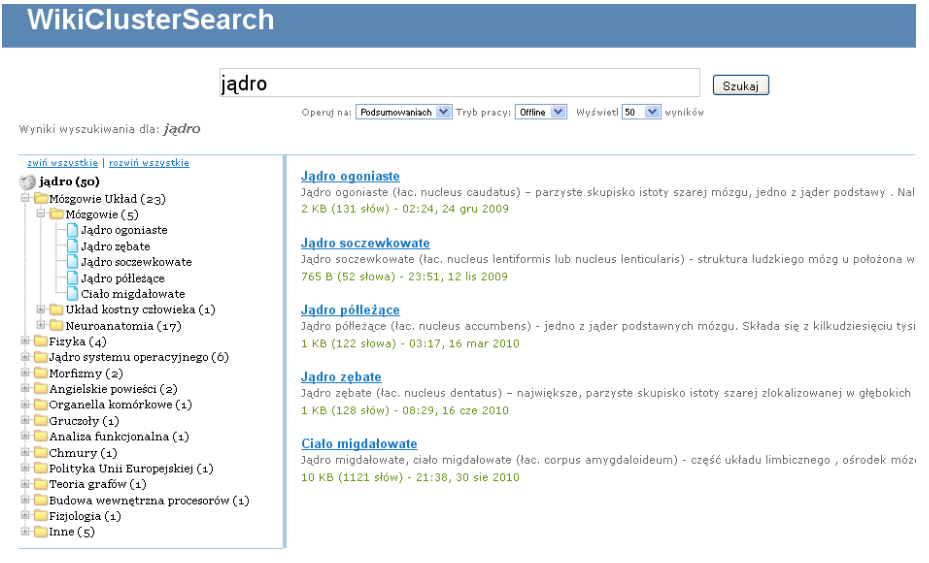


Fig. 7. User interface of the application for clustering Wikipedia search results

We implemented the method presented here in the form of the system that provides clusters for keyword-based search within Polish Wikipedia. The prototype of the system is accessible on-line under url http://swn.eti.pg.gda.pl/UniversalSearch. The screenshot of the application have been presented in Figure 7. The system using the method forms in the fly clusters for the articles that has been selected from Wikipedia using keyword search. In the in Figure 7 we present sample clusters formed for the results returned from Wikipedia for Polish word *jądro* (kernel). In future we plan to implement searching based on clustering for English Wikipedia.

The application shows that forming clusters for Wikipedia pages is useful for organizing the search results. In future we plan to evaluate the clustering results according to the human judgments. Introducing human factor makes this task hard because it

requires reviewing the search results manually and for each cluster decide whether the articles are related to proper cluster correctly or not.

Experiments shown here were performed on a limited set of articles. We plan to perform clustering computations on the whole Wikipedia, to introduce for this source of knowledge new, automated category system. It requires us to take into account some additional issues related to efficiency and requires reimplementation of algorithms to be run on clusters instead of single PC. Create machine-made system of the Wikipedia categories for articles will allow to improve the existing one through finding missing and wrong assignments. Application of this method is also possible for non-categorized documents repository. It allows users to find information using similarity and associations between textual data which is a different approach to the paradigm based on keyword-search.

We plan to research other methods of text representations. We will examine approach to the representation of documents based on algorithmic information [20]. In this approach the similarity between two articles is based on information complexity and is calculated from the size differences of the compressed files [21]. We also plan to research representations based on text semantics. The main idea is to map articles into a proper place of the Semantic Network and then calculate distances between them. As the Semantic Network we plan to use WordNet dictionary [22]. We will use word disambiguation techniques [23] that allow to map words to their proper synsets to perform such a mappings. We made some research in this direction and the first results are very promising [24].

Acknowledgements

The work has been supported by the Polish Ministry of Science and Higher Education under research grant N519 432 338.

References

1. Jolliffe, I.: Principal component analysis. Springer, Heidelberg (2002)
2. Kohonen, T., Somervuo, P.: Self-organizing maps of symbol strings. Neurocomputing 21, 19–30 (1998)
3. Hayes, P., Carbonell, J.: Natural Language Understanding. Encyclopedia of Artificial Intelligence (1987)
4. Allen, J.: Natural language understanding. Benjamin-Cummings Publishing Co., Inc., Redwood City (1995)
5. Russell, S., Norvig, P., Canny, J., Malik, J., Edwards, D.: Artificial intelligence: a modern approach. Prentice-Hall, Englewood Cliffs (1995)
6. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval. Addison-Wesley, Reading (1999)
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34, 1–47 (2002)
8. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18, 613–620 (1975)
9. Kohonen, T.: The self-organizing map. Proceedings of the IEEE 78, 1464–1480 (1990)

10. Gersho, A., Gray, R.M.: Vector quantization and signal compression. Kluwer Academic Pub., Dordrecht (1992)
11. Blachnik, M., Duch, W., Wieczorek, T.: Selection of prototype rules: Context searching via clustering. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 573–582. Springer, Heidelberg (2006)
12. Duch, W., Naud, A.: Multidimensional scaling and Kohonen's self-organizing maps. In: Proceedings of the Second Conference of Neural Networks and their Applications, vol. 1, pp. 138–143
13. Merkl, D.: Text classification with self-organizing maps: Some lessons learned. Neurocomputing 21, 61–77 (1998)
14. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: Websom – self-organizing maps of document collections. In: Proceedings of WSOM, vol. 97, pp. 4–6. Citeseer (1997)
15. Berkhin, P.: A survey of clustering data mining techniques. Grouping Multidimensional Data, 25–71 (2006)
16. Jian, F.: Web text mining based on DBSCAN clustering algorithm. Science Information 1 (2007)
17. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of 2nd International Conference on Knowledge Discovery, pp. 226–231 (1996)
18. Rauber, A., Merkl, D., Dittenbach, M.: The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks 13, 1331–1341 (2002)
19. Koikkalainen, P., Oja, E.: Self-organizing hierarchical feature maps. In: 1990 IJCNN International Joint Conference on Neural Networks, pp. 279–284 (1990)
20. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and its Applications, 3rd edn. Springer, Heidelberg (2008)
21. Bennett, C., Li, M., Ma, B.: Chain letters and evolutionary histories. Scientific American 288, 76–81 (2003)
22. Miller, G.A., Beckitch, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory. Princeton University Press, Princeton (1993)
23. Voorhees, E.: Using WordNet to disambiguate word senses for text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 171–180. ACM, New York (1993)
24. Szymański, J., Mizgier, A., Szopiński, M., Lubomski, P.: Ujednoznacznianie słow przy użyciu słownika WordNet. In: Wydawnictwo Naukowe PG TI 2008, vol. 18, pp. 89–195 (2008)

An Ontology Based Model for Experts Search and Ranking

Mohammed Nazim Uddin, Trong Hai Duong, Keyong-jin Oh, and Geun-Sik Jo

School of Computer and Information Engineering,
Inha University, Korea
{nazim,okjkillo}@eslab.inha.ac.kr, haiduongtrong@gmail.com,
gsjo@inha.ac.kr

Abstract. Experts finding is an important issue for finding potential contributors or expertise in a specific field. In scientific research, researchers often try to find an experts list related to their interest areas to acquire the knowledge about state arts of current research and novices can get benefit to find new ideas for research. In this paper, we proposed an ontological model to find and rank the experts in a particular domain. First, an Academic Knowledge Base(AKB) is built for a particular domain and then an academic social network (ASN) is constructed based on the information provided by the knowledge base for a given topic. In our approach, we proposed a cohesive modeling approach to investigate academic information considering heterogeneous relationship. Our proposed model provides a novel approach to organize and manage the real world academic information in a structural way which can share and reuse by others. Based on this structured academic information an academic social network is built to find the experts for a particular topic. Moreover, the academic social network ranks the experts with a ranking scores depending upon relationships among expert candidates. Finally, we verify the experimental evaluations of our model which improve precision of finding experts compare to baseline methods.

Keywords: Expert search, Ontology, Knowledge base, Academic Social Network.

1 Introduction

The web contain millions of information for a particular keyword or topic provided by many potential contributors and other sources. Finding experts for a specific field from existing web is a challenging task for information seekers. The task of expert finding accomplish by a ranked list of pioneers with relevant topic. Several platforms are introduced to provide experts information such as DBPL[1], CiteSeer[2], and Google Scholar[3]. However, none of these popular platform employed any semantics based information for searching experts. Semantic information provides organization of content in structured way for better understanding and reuse. Additionally, social networks play an

[1] www.dblp.com
[2] www.citeseer.com
[3] www.google.com

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 150–160, 2011.

important role for sharing information in decision making for solving diverse problem. Since academic social network organized information related to academia for various purposes it entails in providing comprehensive services in the academic research field. More important part of academic social network is extraction of relations to expand and model the network for a specific application. In this paper, we have proposed a novel approach to model the academic social network considering the several criteria and explore the relationships for ranking the expert candidates.

There are various approaches that focus on expert finding such as Fine-Grained [12], Hybrid model of topic and language model [12], combine evidence [13]. Additionally, many approaches related to academic social networks have been proposed for finding experts by number of researchers. These approaches include ArnetMiner [6], [9] [8], [10]. However, all of them carry out experts search by considering probabilistic methods and text mining approach. The initial steps of expert finding employed a collection of information describing the skills and knowledge of each individual in a particular domain. And, based on the information provided by each individual a rank list is created according to expertise on related field for a given query. The basic language model measure the association between query term and document written by person for making expert candidate relevant to a query [14]. An improvement of basic language model proposed in [12]. An evidence-oriented probabilistic model for experts search has been proposed in [11] where, relationship between a person and a topic extracted from the specific document. A social network based on profile matching proposed to address the expert finding problem in [7]. The researcher's network limited to matching common interest among researchers and lack of sufficient semantic information to cover the full scientific research domain to find and rank the experts.

We proposed an ontology based model to find and rank the experts in a particular domain, basically research area in computer science and engineering field. Our contribution in this area include following components.

- Build a knowledge base for academic information
- Construct and model an academic social network based on knowledge base for a given topic
- Rank the experts to provide search services for experts related to a specific topic

Building an academic knowledge base is organizing academic information in a structural way using ontology which can be applied to any domain and reuse by other system. Academic information are modeled by meta data analysis and exploring different relationships. Motivation behind this research is to provide an expert search and ranking approach by analyzing the semantics of academic information. Based on semantic academic information an academic social network is modeled considering the relatedness of an expert candidate to a given topic and relationships among neighbors in the network. Finally, ranks the expert candidates by measuring a score for each individual according to expertise in a particular topic area and relationship among other contributors. Content-based evidence and social citation network analysis for finding experts are investigated in [13]. The impact of combining two different sources of experts such as content-based and social networks on expert finding are tested in a average-sized workgroup. Extraction and mining of academic social networks aims at providing comprehensive services in the scientific research field [6]. A social network schema

of FOAF [15] are extended for extracts researcher profiles from the web and proposed three methods for modeling topical aspects of papers, authors, and publication venue. Finally, they construct an academic social network considering author-conference-topic with three different probabilistic approach for finding experts related to topic of a given query. A social network based on mailing list in the W3C corpus [4] is investigated in order to find the expert candidates. According to their experiments, the candidate experts are not well connected in email based social network and performed poor with the HITS algorithm [5].

The rest of the paper is organized in the following way. Section 2 is about the outline of our framework. Section 3 explain the building of academic knowledge base. Construction of academic social network is described in section 4. In section 5 we provide experimental evaluation details and finally conclude our work in section 6.

2 Overview of Expert Finding System

In this section we describe overview of our expert search model. Fig.1 shows the architecture of our system. The system consists of four main components.

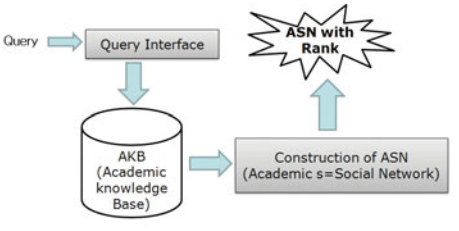


Fig. 1. System Overview

– *Query interface*: A query interface provide an environment for the user to communicate with the system. A query interface might be an interactive GUI with several navigation options or just simple interface provided by the traditional search engine like Google or Yahoo. Users fire a query via interface and get results back related to a given query. In our approach, we just considered the query interface as a simple interactive crossing point to interconnect with the system.

– *Academic Knowledge Base (AKB)*: AKB provide the ontological representation of scientific research in the field of computer science. These information include publications, authors and topics related to the publications and relationships among them are stored in AKB. Details description about AKB present in Section 3.

– *Construction of Academic Social Network (ASN)*: This module describes the logic of constructing the academic social network related to a given query. ANS is created by modeling the information stored in AKB. ANS construction details explain in the Section 4.

3 Building Academic Knowledge Base

We employ the ontological approach to build the AKB for experts finding and ranking. Ontology represents an overview of the domain related to a specific area. In ontology concepts and relationships among concepts are modeled with a high level of abstractions.

3.1 Ontology

Ontology [16] is a formal explicit description of concepts (classes) in a domain with set of properties. Properties of each concept describe the various features and attributes of the concept. Ontologies can be defined as a conceptual graph [2] to represent the queries and resource descriptions.

3.2 Ontology Description of AKB

3.2.1 Classes

Creation of ontology for the scientific research of computer science domain we have defined three top classes listed below.

Researchers: Includes set of all author's information who have contribution in scientific research related to a particular domain. Author's information includes general details of authors like , "Name", "Email address", "Home page", " Affiliation ", "Status" etc.

Publications: Publications class contains the detail about publications of related topics in the domain. Publications class categories into four different subclasses: *Book*, *Journals*, *Proceeding*, and *Technical Reports*, which classify the publications. Publications details include "Title", "Author", "Co-Author", "Keywords", and "Abstract". Particularly, Author will indicate the first author and Co-Authors are other than first author of the publication. Title, Keywords and Abstract are the general format of every scientific paper. In our approach, set of index keywords are the important features of the publications to identify in which topic of the domain the publications are belonging to. If a publication does not supply the keywords then it could be generated by analyzing the abstract.

Fields: Describe the different topic in the domain. In our AKB, we select the computer science domain. So, for the topic we take a reference of ODP^1 (Open Directory Project) hierarchy for selecting topics of the fields. ODP is an open content directory of web pages maintained by a community of volunteer editors. It uses a taxonomic approach to represent topics and web pages that belong to these topics. Part of AKB is shown in the Fig.2.

3.2.2 Relations in AKB

We define four relations which include *has-Publication*, *belong-to-Field*, *written-By*, and *include* in our AKB. As shown in Fig. 3, depicts the relationship between classes. The circle in the figure represents the individual of the class and arcs are the relationship among classes.

A researcher has a publication to a specific topic. For example, a researcher (R) has a publication $(has - publication)$ P,which is related to $(belong - to - field)$ to a

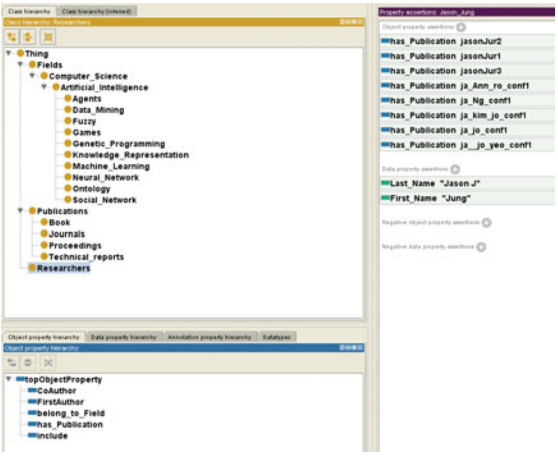


Fig. 2. Part of AKB

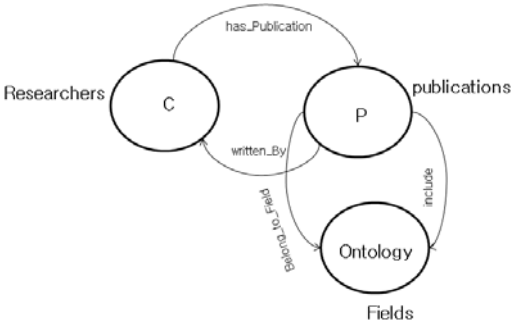


Fig. 3. Relation in Classes

particular topic such as *Ontology* $(Concept)$. Additionally, we have *"include"* relation such as "Fields *include* publications" , by this relation we can find all the publications related to a particular topic under the "Field" class. More clearly, if many researchers have publications related to the concept *Ontology*, we can find all the publications listed under *Ontology* which is sub-concept of the Field class.

3.3 Ontology Population

This section describes how to create instance for the AKB ontology.

- *For Researcher*: Web resources like *DBLP, citeseer* provides list of publications of scientific research. We can extract the researcher's information from the web to create instances for researcher's class.

- *For Publications*: *DBLP* only provides the author name and title of the publications but do not provide keywords and abstract. Beside, *citeseer* provides the publication

content from where we can collect the keywords and abstract. Set of keywords are important feature to summarize the publication content. We have created a set of feature vector for every publication in our knowledge base. Feature vectors are generated by measuring how frequent the index keywords appeared in the content of the publication. The vector space model are applied to generate the feature vectors.

– *For Fields*: Every topic (subclass of Fields) in the "Fields" represented by a set of feature vectors generated from document belong to the topic in the ODP. Documents are basically subsumption and web pages linked to the leaf nodes. Set of feature vectors are generated by *TF-IDF* approach from the ODP documents [1].

Every publication is belong to a particular topic of the "Fields" by the *belong-to* relation. We can assign the publications to related topic automatically by calculating a degree relevancy. Degree relevancy can be measured by the traditional cosine similarity between feature vectors of publication and topic under fields. Details about feature vectors matching described in our previous research [19] . As an example, "Ontology" is a subclass of "Fields" which have the feature vectors $(\overrightarrow{f_1}, \overrightarrow{f_2}, \overrightarrow{f_3}....\overrightarrow{f_n})$ to represents the topic "Ontology" and Similarly, every Publication P has a set of keyword vectors $(\overrightarrow{k}_1, \overrightarrow{k_2}, ..\overrightarrow{k_n})$. Publication P_i is assign to topic T_j by matching the degree relevancy, e.g., a publication P_1 is assigned to topic $(ontology)$ with the degree relevancy of 0.7.

4 Academic Social Networks

Academic Social Network defines as a modeling of information related to academia. These information include authors, publications, topic related to the publications and relationship among them. Several probabilistic methods discussed in Section 1 are introduced by number of researchers and successfully applied for modeling the academic information. Most of the approaches are based on probability measures of text mining and each approach has their own limitation to model the academic information completely. However, we proposed an Academic Social Network (ASN) based on academic knowledge base for modeling scientific research information. The key role players of ASN are the researchers or the authors who have the contributions in scientific research area. Researchers are the candidate to be an expert according to their contributions in a specific field and relationship with other researchers. Experts are determined by measuring scores for each individual based on contributions for a specific topic and relationship with others. In our ASN model, we have employed two phases to evaluate scores for an expert candidate. Initially, we analyze the contributions of each individual candidate based on knowledge base for a given topic and then, update the scores with the relationship with other candidates.

4.1 Construction of Academic Social Network (ASN)

In our approach, an Academic Social Network (ASN) is represented as weighted directed graph $G = (V, E, f_v, f_e)$, where

- V is a set of nodes representing the researchers
- $E \subseteq V \times V$ is a set of edges representing the relationships
- f_v is the node weight function
- f_e is the edge weight function

ASN is constructed based on AKB for a specific topic which is usually a given query within two phases such as Topic-document relationship model and, Author and Co-author relationship model. In the first phase, relationships among a given topic and documents (Publications) are defined according to the methods describe in Section 4.1.1. In the next phase, the ASN is update based on the first author and other authors(co-authors) relationships. Several features are considered when explore the Author and Co-Author relationships model describe in the Section 4.1.2. Basically our intention is to rank the authors with important information and relationships. In our approach, we have investigated the similar concept found in CARRnan [3] but modified it to recognize the potentially important researchers with the realization of following features. An author is imperative if

- He/she has written many papers as a first author
- He/she has included by number of authors (co-author)
- He/she is cited by number of researchers
- He/she has a relation with author who has higher score
- He/she has higher relation weight to any other authors
- A relation weight is higher if it starts from a author of higher score

First three features are conducted in the topic-document relationships model and the rests are model in the second phase which will explore the author and co-author relationships in ASN.

4.1.1 Topic-document Relationship Model (TRM)

For a given topic, topic is matched with the relevant instances of "Fields" class in the AKB. All the publications related to topic are extracted from the knowledge base. An initial score is measured for all the authors (including co-authors) exist in the publications. Suppose, $P_i = (P_1, P_2, P_3...P_n)$ is the set of publications with degree relevancy weight , $W = (w_1, w_2, w_3...w_n)$ in the Fields topic (instance of Fields) of corresponding publication for a given topic. The initial score of a researcher can be calculated by equation 1.

$$P(c|t) = \frac{\alpha \sum_{p \in V_c} w(c|1,p) + \beta \sum_{q \in V_C} w(c|2,p)}{\sum_{i=0}^{n} w_i} \tag{1}$$

Where, c is the expert candidate (researcher/author), t is a given topic, $w(c|1,p)$ relevant degree of publication P as a first author and $w(c|2,p)$ relevant degree as a co-author. α and β are two damping factors where, $\alpha + \beta \leq 1$. Usually first author of the paper has the higher weight then the other authors (co-authors).

4.1.2 Author and Co-Author Relationship Model (ARM)

Social network includes heterogeneous relationships among resources. Different types of relationship imply the different importance for connected resources. In our ASN, we explore two relations based on author and co-authors found in scientific publications to construct academic social network.

Definition 1 (Outward relation)
For any nodes $v_i \in V$ in graph G , the outward relation of v_i are defined as $O_{v_i} = v_j | V_j \in V$, if there exists an $e \in E$ edges between v_i to v_j.
This relation indicates the link between an author(first author) and co-author(s).

Definition 2 (Inward relation)
For any nodes $v_i \in V$ in graph G , the Inward relation of v_i are defined as $I_{v_i} = v_j | V_j \in V$, if there exists an $e \in E$ edges between v_j to v_i.
Inward relations are the links between co-authors and an author.

Considering the features of being a researcher potentially important ASN is constructed based on *Outward* and *Inward* relations. The first three of features defined in Section 4 are considered to calculate the initial scores of expert candidates as described in topic-document relationships model. And, rest of the features are considered when model the ASN. Basically, ASN is constructed by exploring the author and co-author relationship for all the authors found in the set of publications of a related topic. At the initial stage, ASN contains all the authors as an expert candidate with initial score and all the relations (Inward, Outward) exist among expert candidates in the network. Hence, relation's weight are measured by the equation 2.

$$r(x, y) = \frac{w(x)}{\sum_{y_i \in I_y} r(y_i, y)} \tag{2}$$

Where, $r(x, y)$ is the relation weight node x (*expert candidate*) to y (*expert candidate*) and y_i is Inward relation of node y. According to relations weight nodes (expert candidates) weight are updated by the equation 3.

$$w(x) = w(x) + \eta \sum_{y_i \in O_x} w(y_i) * r(x, y_i) + \mu \sum_{y_i \in I_x} w(y_j) * r(y_j, x) \tag{3}$$

Where, O_x is the Outward relation of node x, η and μ is damping factors for Outward and Inward relations.

5 Experimental Evaluations

In this section, we define the experimental details of our proposed model. Expert finding carried out with a given topic from a rich collections of data repositories containing the necessary documents related to topics. Several data repositories such as DBLP, CiteSeer, Google Scholar, Libra, Rexa provide the related data regarding scientific research from which expertise can be derived. However, accessing the data sets of these web repositories is not easy due to the privacy issue.Though DBLP is a good choice to obtain the expert candidates and publications but has some limitations. DBLP does not provide any details about the publication except titles and authors of the publications. Moreover, there are no topic hierarchies exist in DBLP for accessing the publications related to a particular topic. For the purpose of conducting the experiment, we have collected the related data by analyzing a research web site with integration of Google Scholar.

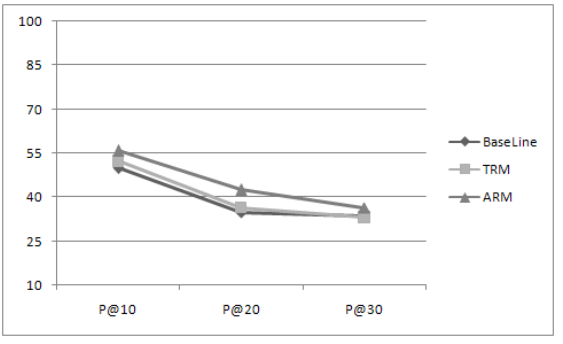


Fig. 4. Evaluation results of our model with the baseline

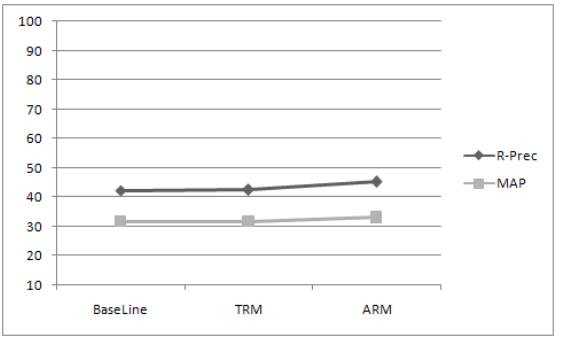


Fig. 5. R-Prec and MAP with the baseline

5.1 Data Collections and Setup

In our proposed model, we have created a knowledge base which includes research information related to computer science domain. Already mentioned earlier, our AKB consists of three main classes and relationships among them. We have collected topics of the Fields class of our AKB with reference of the concept hierarchy ODP of computer science domain. To build the AKB, we have collected the academic research publications from our research lab repositoy [4] along with Google Scholar. IES[4] lab provides all the publications of current researchers and alumni with the abstract. This data set is not sufficiently enough to cover up all the topic of the ”Fields” in the AKB. So we utilize the Google Scholar to collect the related information to build up our AKB. We have made each topic of the ”Fields” class as query to Google Scholar and collected the top 50 records which are most related to the topic. The meta data are collected from the Google Scholar automatically by a crawler program and parser. Finally, we assigned 50 records for each related topic exits in the Fields class of AKB. The processes of assigning all records in AKB are done by manually for the experiment of our proposed model.

[4] http://eslab.inha.ac.kr

5.2 Evaluation Results

To evaluate the experimental results we investigate the method of pooled relevance judgments[6], [12] with the human judgments. Initially for a given query top 50 results were given to some researchers including research faculty members, doctoral and master students to assess the expert candidates returned by our system. To help the researchers in evaluation process we have provided necessary information of expert candidates to them. Assessments were carried out mainly considering the features described in subsection 4.1 to become an author potentially important. When performed the experiments the value of η and μ were set 0.7 and 0.3 respectively.

In our experiments, we conducted evaluation of our proposed model in terms of P@10, P@20, P@30, R-Prec and MAP [18],[12][6]. P@n denotes the precision at the rank n with respect to n retrieved results for a given query. R-precision(R-Prec) is defined as the precision after R documents are retrieved where R is the the number of relevant documents retrieved for a given topic. Mean Average Precision (MAP) is the mean of precision values obtained after each relevant documents retrieved with precision of relevant document which are not retrieved using zero. Using these metrics fig. 4 and 5 shows the evaluation results our model with a baseline scores. There are no significant differences between baseline methods and our TRM model but in the ARM, a considerable improvement is observed with the baseline model.

6 Conclusions

In this paper, we proposed an ontological model to find experts for a particular domain. An Academic Knowledge Base (AKB) is built with publication details of authors to accomplished expert search for a given query. Additionally, an Academic Social Network (ASN) is built by analyzing the author, co-author relationships to boost the expert search with ranking scores. Finally, ASN provides all the expert candidates and their relationship among other candidates with rank. Specially, we employed the ontological approach to model the academic details for a particular domain. The model describes a structural way to maintain the academic information for finding the experts. The main advantage of such semantic approach is it can be reuse and shared by other system. Assignment of meta data to the ontology is made manually in this approach. Our future task is to complete the knowledge base with automatically assigning meta data from a rich data set. Additionally, we will apply some inference analysis to check the consistency of AKB.

References

1. Duong, T.H., Mohammed, N.U., Jo, G.S.: A Collaborative Ontology-Based User Profiles System. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 540–552. Springer, Heidelberg (2009)
2. Guarino, N., Masolo, C., Vetere, G.: OntoSeek: Content-Based Access to the Web. IEEE, Intelligent Systems 14(3), 70–80 (1999)

3. Wu, G., Li, J., Feng, L., Wang, K.: Identifying potentially important concepts and relations in an ontology. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 33–49. Springer, Heidelberg (2008)
4. Chen, H., Shen, H., Xiong, J., Tan, S., Cheng, X.: Social Network Structure behind the Mailing Lists. ICT-IIIS at TREC Expert Finding Track; TREC06, working notes (November 2006)
5. Jon, M.K.: Authoritative Sources in a Hyperlinked Environment. ICT-IIIS at TREC Expert Finding Track. TREC, working notes (November 2006)
6. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Zhong, S.M.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: KDD, Las Vegus, Nevada, USA (August 2008)
7. Duong, T.H., Nguyen, N.T., Jo, G.S.: Constructing and mining a semantic-based academic social network. Journal of Intelligent & Fuzzy Systems 21(3), 197–207 (2010)
8. Harrison, T.M., Stephen, T.D.: The Electronic Journal as the Heart of an Online Scholarly Community. Library Trends 43(4) (Spring 1995)
9. Newman, M.E.J.: The structure of scientific collaboration networks. In: PNAS, vol. 8(2), pp. 404–409 (2001)
10. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. In: PNAS, vol. 101, pp. 5200–5205 (2004)
11. Bao, S., Duan, H., Zhou, Q., Xiong, M., Cao, Y., Yu, Y.: A Probabilistic Model for Fine-Grained Expert Search. In: HLT, pp. 914-922. ACL (2008)
12. Deng, H., King, I., Michael, R.L.: Formal Models for Expert Finding on DBLP Bibliography Data. In: Eight IEEE Internation Conference on Data Mining, pp. 163–172 (2008)
13. Bogers, T., Kox, K., Bosch, A.: Using Citation Analysis for Finding Experts in Workgroups. DIR, Maastricht, the Netherlands, pp. 14–15 (April 2008)
14. Krisztian, B., Leif, A., de Maarten, R.: Formal Models for Expert Finding in Enterprise Corpora. In: SIGIR, ACM, New York (2006), ISBN: 1-59593-369-7
15. Brickley, D., Miller, L.: Foaf vocabulary specification. Namespace Document (September 2004), http://xmlns.com/foaf/0.1/
16. Natalya, F.N., Deborah McGuinness, L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, Stanford, CA, 94305
17. Steyvers, M., Smyth, P., Michal, R., Griffiths, T.: Probabilistic Author Topic Models for Information Discovery. In: Proc. of SIGKDD (2004)
18. Buckley, C., Voorhees, E.M.: Retrieval Evaluation with Incomplete Information. In: Proc. of SIGIR, vol. 4, pp. 25–32 (2004)
19. Mohammed, N.U., Duong, T.H., Jo, G.S.: Contextual Information Search Based on Ontological User Profile. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS, vol. 6422, pp. 490–500. Springer, Heidelberg (2010)

A Block-Structured Model for Source Code Retrieval

Sheng-Kuei Hsu[1,2] and Shi-Jen Lin[1]

[1] National Central University, Jhongli 320, Taiwan, ROC
[2] Nanya Institute of Technology, Jhongli 320, Taiwan, ROC
skhsu@nanya.edu.tw, sjlin@mgt.ncu.edu.tw

Abstract. The large amounts of software source code projects available on the Internet or within companies are creating new information retrieval challenges. Present-day source code search engines such as Google Code Search tend to treat source code as pure text, as they do with Web pages. However, source code files differ from Web pages or pure text files in that each file must follow a set of rules called syntax, and a source file can be seen as a structured document. Each file contains elements to complete a task. In this paper, we parse each source code file into elements called blocks. They include a non-leaf block and a leaf block for further indexing and ranking. These leaf blocks can be categorized into code-data and meta-data blocks that possess different stemming and stop-word filtering processes used in building the source code index. Finally, to provide a flexible code search scheme, we also propose a block-specified query scheme. Experimental results indicate that our approach provides a more flexible code search mechanism that results in a higher number of relevant items.

Keywords: code search, code retrieval, block structured model, vector space model.

1 Introduction

The growth of source code projects available on the Internet or within companies is creating new challenges in the field of information retrieval. Although these large amounts of source code result in new search challenges, they also yield new search opportunities that contribute to an improved understanding of software engineering data, including requirements, designs, and source codes. Searching such software repositories is important in developing a thorough understanding of software requirements, design, and evolution, as well as in developing ways to reuse these software components. Source code information, which is more structured than simple text documents, can be easily categorized into blocks or fields, such as field declarations, methods, or comments. Such an organization readily supports specialized searches. The field declaration statements and comments related to each source file are preserved as blocks that enable searches by means of specifying block names.

Web search engines have become one of today's most useful tools; accordingly, numerous search models have been developed. In recent years, these traditional IR approaches have been applied to code searching, such as Google Code Search [1] and

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 161–170, 2011.

Korders.com [2]. However, various problems stem from the differences between Web documents and source files. In our research, we see source files as a hierarchy of blocks and categorize them into two types of blocks: non-leaf blocks and leaf blocks. A non-leaf block comprises several leaf blocks, and a leaf block can be further categorized into code-data and meta-data blocks. A code-data block, such as a field-declaration block, contains code statements that always reuse the third-party Application Programming Interface (API) or previous components developed by the same team. A meta-block such as a comment block contains the meaning that developers have given a code-block or lines of code.

Recent source code search engines such as Google Code Search tend to treat source code as pure text, as they do with Web pages. However, as source code files differ from Web pages or pure text, when applying IR approaches to code search, certain issues arise in the indexing and ranking stages. The first challenge we face when studying source code retrieval is determining what the ideal indexing and ranking unit will be for an application developer. In one source code file, there are many possibilities. We can return a whole source file or any of its sub-elements, but what is the best retrieval unit? The answer depends on the user query and the content of each element. Second, terms within code-data blocks (code statements) are processed in the indexing stage by means of a stemming process, which leads to low precision. Thus, searching for "connection" will retrieve "connected," "connecting," "connection," and "connections," even though the user only wishes to search through the API-class name "connection." Finally, API-terms are incorrectly filtered out. For example, "author" may be an API-class name, but it will be filtered out because it is included on a common English stop-word list. On the other hand, the API "string" should be filtered but is not. In the scoring stage, the traditional term-weighting approach regards terms within a meta-area as title words and weights them. If a term appears in both the code-data block and the meta-data block, those files that contain only an API term will not be found. For example, in a Java program, "connection" can be an API-class name, a file name (such as "connection.java"), or merely a common term within a comment. With a TF-IDF approach, a term within a file name or within a comment block will be assigned greater weight than general terms; this led to the inability to find API usage examples of "connection."

To deal with these problems, first, we propose a block-structured model to segment a source code file into a hierarchy of blocks. These blocks include file, class, method, and code blocks and are categorized into code-data and meta-data blocks that possess different stemming and stop-word filtering processes used in building the source code index. At this stage, we skip the stemming process and propose a new stop-word list for the code-data blocks to filter out those frequent "code words." Second, we segment source files into a hierarchy of blocks. This provides a flexible and ideal source code unit for further indexing and ranking processes. This can return to users an ideal retrieval unit to solve information needs. Finally, we also offer a block-specified query to assist those users who may only query for specific blocks in the source code repository.

2 Related Work

Matching keywords is the most prevalent approach. Early work in this area demonstrated that keywords from comments and variable names were often sufficient for finding reusable routines [3, 4]. Later work focused on query refinement either directly or by looking at what the programmer was doing and making use of an appropriate ontology, learning techniques, natural language, collaborative feedback, or Web interfaces [1, 2, 5]. Keyword-based searches typically yield numerous unevaluated results. The user must read each instance of returned code and attempt to understand these search results. However, due to their simplicity and ease of use, keyword-based searches remain popular.

The second approach involves finding with structure. It uses information about a source code file structure [6, 7, 8], such as method signatures, return types, or expected loop statements. Stratchoma [9] uses structured context to automatically formulate a query to retrieve code samples with similar contexts from a repository. The results are ranked based on the highest number of structured relations contained in the results. Sourcerer [10] is a recent study proposed to automate crawling, parsing, fingerprinting, and database storage of open-source software on an Internet-scale. In addition to keyword searching, it provides a novel search method—fingerprints.

Using test cases to specify what to search for comprises the third approach. In extreme programming and test-driven software development, first, the developer writes a failing automated test case that defines a desired improvement or new function, then produces code to pass that test. Treating a test case as a query, CodeGenie [11] is a tool that implements a test-driven approach to search and reuse source code available in large-scale code repositories. The search approach still poses two problems: 1) it is difficult to write the test cases and 2) the solution might be not well-defined.

Recent commercial work, such as Google Code Search, incorporates traditional information retrieval approach and is applying Web search engine capabilities to code searching. Such keyword-based code search engines treat source code as pure text, as they do with Web pages. However, as source code files differ from Web pages or pure text files, when applying IR approaches to code searching, certain issues arose. Our approach is similar to that of Google Code Search or Korders, which continue to utilize keyword searches, but it incorporates a structured block approach to indexing and searching.

3 Architecture

Figure 1 shows the architecture of our code search engine. The arrows show the flow of information between various entities or processes. Information on each entity or process is listed below.

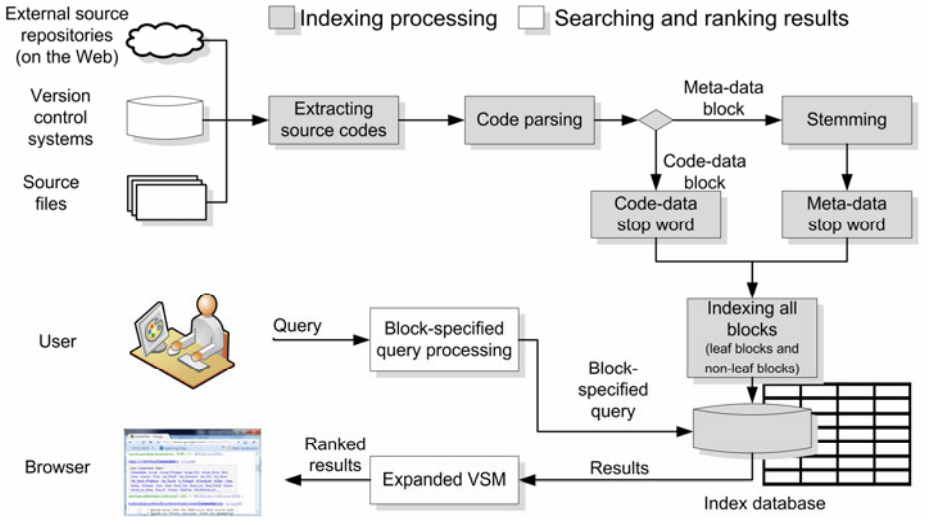


Fig. 1. The architecture of our code search engine

- *External code repositories*: These are the source code repositories available on the Internet—such as the well-known repository Sourceforge.com—or within a company. They include source code repositories on the Web, in version control system, or in source code files.
- *Extracting source code*: The process is to retrieve source code files from previous external code repositories. We retrieved these source code files manually.
- *Code parsing*: We developed a specialized parser to parse every source code block (or field) from a source code project. The parser can identify each type of block, such as field-declaration or comment, and extract keywords (or terms) from each block (Figure 2). In addition, the process divides the content of each file into two types of blocks—meta-data and code-data blocks—for the following stemming and stop-word process.
- *Stemming, stop words, and indexing*: For each meta-data block, the process uses stemming and meta-data stop-word processes, as traditional information retrieval does. For each code-data block, the process skips the stemming step to keep the original word. For example, we do not need to identify the string "Connection" as based on the root "Connect" in a code-data block. In addition, we developed a special stop-word list, called code-data stop words for code-data blocks. Finally, the process indexes all components as an information unit (block), including file, class, method, and field declaration. A higher-level unit includes the contents of sub-elements.
- *Querying*: In this process, a user can provide keywords, as he or she does on Google, or assign a specified block (field) for search, such as the comment block or field-declaration block.

- *Ranking*: Once a query has been created and submitted to the ranking process, the process measures the similarity between the query and all possible combinations of blocks. In the study, a common similarity measure, known as the cosine measure, was adopted to rank the source files. Unlike traditional IR regarding a document (file) as a scored entity, we regard a block as an entity.

4 A Structured Model for Indexing and Ranking

4.1 Multiple Blocks for Indexing

In our approach, we index all elements (blocks) in source code files. First, each source file is thought to consist of distinct parts, or blocks, such as file name, class name, import declaration, code block, and so on (Figure 2). For different processing, we divide these blocks into non-leaf and leaf blocks.

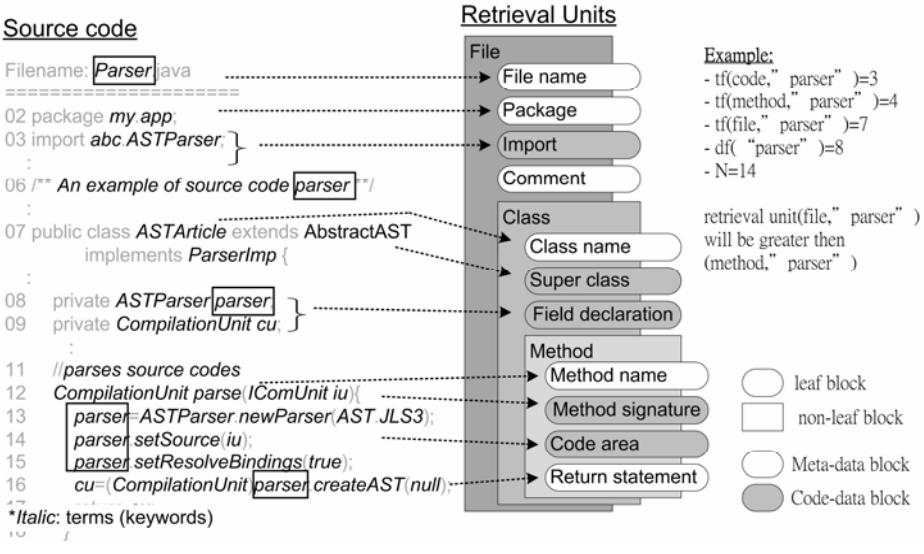


Fig. 2. Identifying blocks within a source code file

- Non-leaf and leaf blocks: In our study, there are three non-leaf blocks: the file, class, and method blocks. Each non-leaf block may include one or more leaf blocks such as the field-declaration, code-area, and return statement leaf blocks. In contrast, each leaf block comprises concrete source code statements and does not include any sub-blocks. Moreover, leaf blocks can be categorized into meta-data and code-data blocks for further processing.
- Meta-data blocks: These blocks contain information about a source file or a block of code. They include file name, class name, package name, and comment. The text inside these blocks will be processed as it is in traditional text information retrieval.

- Code-data block: These blocks are composed of code statements, including import declaration, field declaration, method name, code block, super class type (the class extended), and interface type (the class implemented). They contain pure code statements that need to be processed with different processes in meta-data blocks.

In the indexing step, the text within each block is tokenized, and terms or keywords are extracted. For each term in a block, including non-leaf blocks and leaf blocks (meta-data blocks and code-data blocks), the number of the source blocks in which it occurs (file frequency) and the IDs of the source files in which it occurs are stored. The inverted index for the set of source files is also stored. In our approach to evaluate all possible combinations of blocks for each query, we index all components as information units.

4.2 Indexing Source Code

In our approach, we use an indexing process for the text. For meta-data blocks, the process is used to tokenize text, extract relevant words, discard common words (common stop-words), and stem the words (reduce them to the root form; for example, "connection," "connections," and "connecting" are reduced to "connect"). For code-data blocks, the process is used to tokenize code data, extract relevant code-data terms, and discard common code-data terms (code stop-words). In indexing code-data blocks, we do not stem the code-data terms to maintain the original forms and thereby allow for more exact code-data searching. For example, when searching for the API term "connection," the term "connect" should not appear in the search results.

4.3 Block-Specified Query

Our structured approach offers a block-specified query to assist those users who may only query for specific blocks in the source code repository. Consider a source code repository R that is accessible by means of a query interface. We define this repository as a collection of source code files.

$$\begin{aligned} R &= \{F_1, F_2, ...F_i, ..., F_n\} \\ F_i &= \{(B_1, T_1), (B_2, T_2), ...(B_k, T_k)\} \\ &\text{where } T_k = \varnothing \text{ or } T_k =< t_1, t_2, ..., t_j > \end{aligned} \tag{1}$$

The source file F_i consists of a set of block names (B_k), and each source file has at least one data block made up of a block name B_k and a block value T_k pair. If a source file contains the block, it has at least one term. Given a source file F_i, for example, consisting of three blocks—class name, comment, and field declaration—with the values of <"Parser">, <"Parse," "Source Code">, and <"ASTParser">, respectively, the source file can be expressed as F_i = {(class, <"Parser">), (comment, <"Parse," "Source Code">), (attribute, <"ASTParser">)}.

A user query is divided into two parts: a block-specified query (BQ) and a block-unspecified query (UQ). We define a block-unspecified query as a set of keywords (or terms):

$$UQ = \{t_1, t_2, ..., t_k\}$$

A block-specified query, BQ, is defined as a set of pairs:

$$BQ = \{B_1 : v_{1q}, ..., B_n : v_{nq}\}, n_q \geq 1,$$

where each B_k is an attribute block, and each v_{kq} is a value with the domain of B_i. Block value v_{kq} contains at least one term.

$$v_{kq} =< t_1, t_2, ...t_i > \ i \geq 1$$

In this research, if a block-specified query was given by a user as BQ = {comment: <"Parse," "Source Code">), (field: <"ASTParser">)}, then the *blocks* correspond to the *comment* and *field-declaration* block.

4.4 Adopting VSM for Ranking Results

We used our expended vector space model to judge the relevance of a block (B) and a query. A block may contain several sub-blocks (b). Then we calculated the similarity between an individual block (B) within a source file and the user query (q) as follows:

$$Sim(B,q) = \frac{B \bullet q}{|B| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,b} \times w_{i,\mathrm{q}}}{\sqrt{\sum_{i=1}^{t} w_{i,b}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2}$$

where, i is a term in the block of a source file; $w_{i,q}$ is the weight of the query, term i does not affect the ranking and is obtained by $w_{i,q} = idf_q$; and $w_{i,b}$ is the weight of the term i within b and is obtained by $w_{i,b} = tf_{i,b} * idf_i$.

To adapt VSM to our source code files, we introduce changes that express their characteristics as follows. First, we consider each block as a new retrieval unit. tf_i becomes $tf_{i,b}$, so:

- $tf_{i,b}$ is the number of occurrences of term i in block b;
- df_i is the number of blocks that contain term i; and
- N is the total number of blocks in the source code repository.

At this point, we have already defined how to calculate the similarity between a block and user query through our adopted VSM approach. We will now introduce a nesting factor, (S). This factor will reduce the term contribution for distant blocks in the structure of a source file. We define our expended VSM as follows:

$$Sim(B,q) = \frac{B \bullet q}{|B| \times |q|} = \frac{\sum_{i=1}^{t} w_{i,b} \times w_{i,\mathrm{q}} \times S_{i,b}}{\sqrt{\sum_{i=1}^{t} w_{i,b}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{3}$$

The nesting factor can vary between the following two values:

- $S_{i,b} = 1$, for terms directly in block b,
- $S_{i,b} = 1/nd$, nd being the depth of the structure.

5 Implementation

In this study, we propose using Lucene, the Java-based open-source search engine, to search source code by extracting and indexing relevant source code elements. The search is restricted to Java source code. However, extending the search to any other source code in any other programming language should not affect the results.

Lucene is one of the most popular open-source search engine libraries. In our research, Lucene was modified to allow the source code to be indexed and searched. Given a set of source code files, the modified system can create indexes and allow users to search those indexes with complex queries such as +comment:Parser -field:ASTParser. (Table 1)

Table 1. Example of querying source code

Query Expression	Matches Source Files that …
Super:JFrame code:JButton	Extends class JFrame and uses JButton class in the code block (method block).
code:ASTParser +import:org.eclipse	Uses ASTParser in the code and definitely has org.eclipse in the import-declaration block.
Field:Document	Uses Document as an attribute in the class.
m:Parser	Contains the term Parser in meta-blocks, including file_name, package_name, comment, class_name, method_name, and return statement blocks.
c:JGraph	Contains the JGraph class in code-data blocks.
method:paint +signature:Graphics	Contains a method named paint and definitely has Graphics in the method signature.
JFrame JButton	Extends class JFrame and uses JButton in the field-declaration block (the block-unspecified query will be translated as, for example, super:JFrame field:JButton).

Once created, indexes can be searched by providing complex queries that specify the field and the term being sought. For example, the user query comment:Parser AND field:ASTParser will result in all source files that contain Parser in the comment block and a field declaration that includes the term ASTParser. Source files that match the query are ranked according to the similarity between the query and the source files stored in the repository.

6 Experiments

The purpose of our experiment was to demonstrate the flexibility and superiority of the proposed structured block scheme over other source code search engines: 1) block-specified queries specified by the user for code retrieval (BQ) vs. Google Code Search

(Google) and 2) block-unspecified queries for code retrieval (UQ) vs. Google Code Search (Google). Two types exist in UQ; V1 refers to cases in which only one best query is used for retrieval, whereas V1-3 describes cases that use the best three block-specified queries for retrieval.

To evaluate the results, the most common experimental measures were taken using IR methods, namely *recall*, *precision*, and *F1* measures. F1 measures combine precision and recall with equal weight and are defined as F1=2PR / (P+R). Moreover, we evaluated these measures on eight open-source projects. These projects were chosen to cover a wide range of applications on Sourceforge.net, including jfroum, ctypes, CoFFEE, Vocabulary Learning Tool, jMusic, jscl-meditor, and jmlspecs. Meanwhile, to ensure a fair comparison with Google Code Search, we limit the Google search to only the 8 projects.

Table 2. Results of the experiments

Scheme/ Measures	Google	BQ	UQ_V1	UQ_V1-3
Precision	45.05%	60.08%	50.12%	54.37%
Recall	60.22%	68.55%	70.29%	75.62%
F1	51.54	64.04	58.52	63.26

As illustrated in Table 2, BQ is superior to UQ and Google. This phenomenon is simply due to the fact that the queries are more suitable for searching source code data. In addition, BQ was constructed by the actual developers and, thus, more thoroughly reflects their information needs than other source code search schemes, including Google, UQ_V1, and UQ_V1-3. The values of the results from UQ_V1 and UQ_V1-3 are similar because the block-specified queries in UQ_V1 are almost identical to the UQ_V1-3 queries.

7 Conclusion

Current commercial code search engines such as Google Code Search incorporate traditional information retrieval approaches while bringing the capabilities of Web search engines into the realm of code searching. However, such keyword-based code search engines treat source code like pure text, just as they do with Web pages, which creates certain issues. To address these issues, we propose a new technique for code searching that utilizes a block-structured scheme with three distinguishing features. First, we incorporated a multi-block indexing method. By treating source code files as multi-block documents, separating them into two types of blocks, and then creating an index database with different processes, we improved the precision of source code searching. Second, we presented a flexible query method by specifying a block. Finally, we provided a flexible retrieval block as a unit of search result to fulfill the user information needs. In comparing this search engine with similar ones, the experiments indicate that the proposed scheme has the capacity to result in more relevant items than other related works because it not only incorporates a new index method but also includes a flexible inquiry index database.

References

1. Google Code Search, http://www.google.com/codesearch
2. Korders.com, http://www.koders.com/
3. Frakes, W.B., Pole, T.P.: An empirical study of representation methods for reusable software components. IEEE Trans. on Software Engineering 20(8), 617–630 (1994)
4. Maarek, Y.S., Berry, D.M., Kaiser, G.E.: An information retrieval approach for automatically constructing software libraries. IEEE Trans. on Software Engineering 17(8), 800–813 (1991)
5. Krugle web site, http://www.krugle.com
6. Bajracharya, S., Ngo, T., Linstead, E., Dou, Y., Rigor, P., Baldi, P., Lopes, C.: Sourcerer: a search engine for open source code supporting structure-based search. In: Proc. OOPSLA 2006, pp. 682–682 (2006)
7. Begel, A.: Codifier: a programmer-centric search user interface. In: Workshop on Human-Computer Interaction and Information Retrieval (2007)
8. Hoffmann, R., Fogarty, J.: Assieme: finding and leveraging implicit references in a web search interface for programmers. In: Proc. UIST 2007, pp. 13–22 (2007)
9. Holmes, R., Murphy, G.C.: Using structured context to recommend source code examples. In: Proc. ICSE, The 27th International Conference on Software Engineering, ICSE 2005, pp. 117–125. ACM Press, New York (2005)
10. Linstead, E., Bajracharya, S., Ngo, T., Rigor, P., Lopes, C., Baldi, P.: Sourcerer: mining and searching internet-scale software repositories. Data Min. Knowl. Discov. 18(2), 300–336 (2009)
11. Lemos, O.A.L., Bajracharya, S.K., Ossher, J., Morla, R.S., Masiero, P.C., Baldi, P., Lopes, C.V.: Codegenie: using test-cases to search and reuse source code. In: Proc. ASE. The twenty-second IEEE/ACM International Conference on Automated Software Engineering (ASE 2007), pp. 525–526. ACM Press, New York (2007)

Identifying Disease Diagnosis Factors by Proximity-Based Mining of Medical Texts

Rey-Long Liu[1], Shu-Yu Tung[2], and Yun-Ling Lu[3]

[1] Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan
rlliutcu@mail.tcu.edu.tw
[2] Winbond Electronics Corporation, HsinChu, Taiwan
sytung@gmail.com
[3] joanna.lu@gmail.com

Abstract. Diagnosis of diseases requires a large amount of discriminating diagnosis factors, including the risk factors, symptoms, and signs of the diseases, as well as the examinations and tests to detect the signs of the diseases. Relationships between individual diseases and the discriminating diagnosis factors may thus form a diagnosis knowledge map, which may even evolve when new medical findings are produced. However, manual construction and maintenance of a diagnosis knowledge map are both costly and difficult, and state-of-the-art text mining techniques have difficulties in identifying the diagnosis factors from medical texts. In this paper, we present a novel text mining technique PDFI (Proximity-based Diagnosis Factors Identifier) that improves various kinds of identification techniques by encoding *term proximity contexts* to them. Empirical evaluation is conducted on a broad range of diseases that have texts describing their symptoms and diagnosis in MedlinePlus, which aims at providing reliable and up-to-date healthcare information for diseases. The results show that PDFI significantly improves a state-of-the-art identifier in ranking candidate diagnosis factors for the diseases. The contribution is of practical significance in developing an intelligent system to provide disease diagnosis support to healthcare consumers and professionals.

1 Introduction

Clinical diagnosis of diseases relies on critical diagnosis factors, including the risk factors, symptoms, and signs of the diseases, as well as the physical examinations and lab tests to detect the signs of the diseases. The diagnosis factors should be capable of *discriminating* individual diseases. As illustrated in Figure 1, the relationships between the individual diseases and their diagnosis factors may form a *diagnosis knowledge map*, which is the key knowledge for healthcare professionals to conduct clinical diagnosis, as well as for healthcare consumers to pinpoint the possible disorders in their bodies. For the purpose of disease diagnosis, we often prefer putting into the diagnosis knowledge map those diagnosis factors that are more capable of discriminating the diseases. A diagnosis knowledge map may also *evolve* as new medical findings are produced in medical research.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 171–180, 2011.

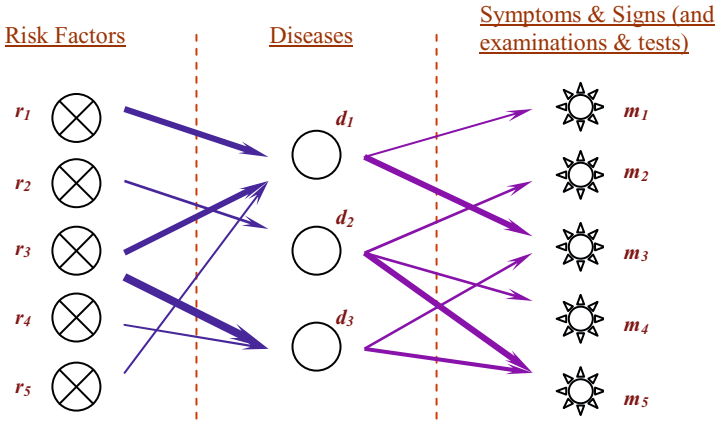


Fig. 1. A disease is caused by its related risk factors and leads to symptoms and signs to be detected by physical examinations and lab tests, and hence a *diagnosis knowledge map* should consist of *many-to-many* relationships between individual diseases and the risk factors, symptoms, signs, examinations, and tests that have different capability of *discriminating* the diseases (indicated by different styles of arrows) and may *evolve* over time

1.1 Problem Definition and Motivation

In this paper, we explore how the identification of the diagnosis factors may be supported text mining systems. The research is motivated by the difficulty and cost of manually constructing and maintaining a diagnosis knowledge map that incorporates a large number of diseases and diagnosis factors, which are both discriminative and evolvable. In clinical practice, a diagnosis knowledge map should consist of a large number of many-to-many relationships with different capabilities of discriminating diseases. However, manual construction of even a single disease may require lots of time and effort (e.g., one person month in a similar attempt noted in [10]), and the complexity dramatically increases when a broad range of diseases are considered and the diagnosis factors may evolve as new medical findings are produced. By identifying candidate diagnosis factors from up-to-date medical texts, a text mining system may support to reduce the difficulty and cost.

Technically, we present a novel technique PDFI (Proximity-based Diagnosis Factors Identifier) that employs *term proximity contexts* to improve various kinds of text-based discriminative factors identifiers. For a candidate diagnosis factor u, PDFI measures how other candidate diagnosis factors appear in the areas near to u in the medical texts, and then encodes the term proximity context into the discriminating capability of u measured by the underlying discriminative factors identifiers. The idea is based on the observation that in a medical text talking about the diagnosis of a disease, the diagnosis factors often appear in a nearby area of the text, and hence by encoding term proximity contexts into the discriminating capabilities of the diagnosis factors, various kinds of discriminative factor identifiers may be improved.

1.2 Major Challenges and Related Work

Major challenges of the research include (1) estimating the discriminating capabilities of candidate diagnosis factors, (2) recognizing term proximity contexts of the factors, and (3) measuring the strengths of the factors by integrating their term proximity contexts and discriminating capabilities. No previous studies tackled all the challenges.

For the first challenge, previous studies on the correlations between biomedical objects (e.g., proteins, genes, and diseases) often employed sentence parsing [9][13], template matching [2][8], or integrating machine learning and parsing [5] to extract the relationships between biomedical objects. The previous techniques are not applicable to the identification of diagnosis factors, since the relationships between diseases and their diagnosis factors are seldom explicitly expressed in individual sentences. Most sentences in a medical text do not repeatedly mention the diseases, nor do they mention "diagnosis" as the relationship between the diseases and the factors.

Another approach to the identification of diagnosis factors is text classification. The problem of identifying diagnosis factors for diseases may be treated as a *feature selection* problem in which we score and select those features (terms of diagnosis factors) that are capable of discriminating the categories (diseases). The selected features may also serve as the basis on which a text classifier may be built to classify symptom descriptions into disease categories [6]. Feature selection in text classification does not rely on processing individual sentences, but instead relies on the texts labeled with category names to measure how a feature may discriminate the categories, making it a plausible way to identify diagnosis factors. Previous studies had developed and tested many feature selection techniques (e.g., [7][14]).

However, the feature selection techniques do not consider proximity contexts of the diagnosis factor terms (features), which is the second challenge of the paper. As noted above, the authors of a medical text for a disease tend to describe the diagnosis factors of the disease (if any) in a nearby area of the text, making term proximity contexts quite helpful to identify the factors. Previous studies noted term proximity as important information as well, however they often employed term proximity to improve the ranking of the texts retrieved for a query [1][3][11][15], rather than the identification of diagnosis factors.

We refer the feature selection techniques as the underlying discriminative factors identifiers, and explore how term proximity may be used to improve them. Therefore, the final strength of each diagnosis factor should be based on both the term proximity context and the discriminating capability of the factor. The estimation of the strength is the third technical challenge, which was not explored in previous studies.

1.3 Organization and Contributions of the Paper

Section 2 presents PDFI, which is an enhancer to various kinds of discriminative factors identifiers. To empirically evaluate PDFI, section 3 reports a case study in which real-world medical texts for a broad range of diseases are tested. PDFI significantly improves a state-of-the-art identifier in ranking candidate diagnosis factors for the diseases. The contribution is of practical significance in the development of an

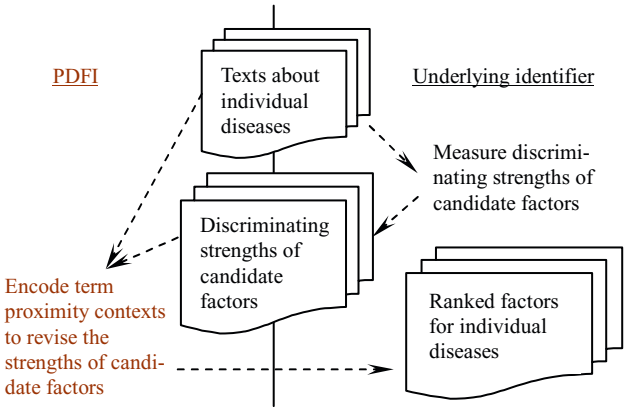


Fig. 2. PDFI employs term proximity contexts to enhance various kinds of discriminative concepts identifiers

intelligent system used in clinical decision support and medical education for healthcare professionals, as well as health education for healthcare consumers.

2 Enhancing Factor Identifiers by Term Proximity Contexts

As illusrated in Figure 2, PDFI collaborates with the underyling discriminative factors identifier by integrating the output of the identifier (i.e., discriminating strengths of candidate factors) with term proximity information, and the resulting strengths of the factors are used to rank the factors for medical experts to validate. The strength estimation is thus essential in reducing the load incurred to the experts.

More specifically, Table 1 presents the algorithm of PDFI in estimating the strength of a candidate factor. Given a feature u (candidate factor term) and its discriminating strength x for a disease c (produced by the underlying discriminative factors identifier), PDFI measures a term proximity score for u (*ProximityScore*, ref. Step 4 and Step 5). The proximity score is measured based on the shortest distances (in the given medical texts) between u and other candidate features of c (ref. Step 4.1 and Step 4.2). PDFI employs a *sigmoid function* to transform the shortest distances into the proximity score–the shortest distance between u and another candidate feature n is transformed into a weight between 0 and 1, and the smaller the distance is, the larger the weight will be (ref. Step 4.3). The sigmoid weighting function has a parameter α that governs the distance for which a weight of 0.5 is produced (when the distance is equal to α, the sigmoid weight is 0.5), and we set α to 30. Moreover, it is interesting to note that, to measure the proximity score of u, the sum of the sigmoid weights with respect to all the other candidate features is computed (i.e., *RawScore*, ref. Step 4.3) and normalized by the maximum possible sum of the weights (i.e., |N|, ref. Step 5). The normalization aims at measuring the *contextual completeness* of the

Table 1. Proximity-based revision of the strength of a candidate factor

Procedure: *ProximityStrength*(u,c,x)

Given: u is a candidate feature (a candidate factor term) for disease c, and x is the preliminary strength of u with respect to c (x is produced by the underlying factor identifier).

Effect: Revising x based on the term proximity context of u in the texts about c.

Method:

(1) $RawScore \leftarrow 0$;
(2) $N \leftarrow \{n \mid n$ is a candidate feature for c , $n \neq u\}$;
(3) $D_c \leftarrow$ Set of texts about c;
// Estimate term proximity score of u
(4) For each feature $n \in N$, do
  (4.1) If u and n do not co-occur in any text in D_c, $MinDist \leftarrow$ Length of the largest text in D_c;
  (4.2) Else $MinDist \leftarrow$ Minimum distance between u and n in the texts in D_c;
  (4.3) $RawScore \leftarrow RawScore + \frac{1}{1+e^{-(\alpha - MinDist)}}$;
(5) $ProximityScore \leftarrow RawScore / |N|$;
// Integrate the term proximity score and the preliminary strength of u
(6) $RankScore(u,c) \leftarrow 1 +$ Rank of u among all candidate features of c (ranked by the preliminary strengths of the features);
(7) $x \leftarrow \frac{1}{RankScore(u,c)} + ProximityScore$;
(8) Return x;

other candidate features appearing near to u, and hence the proximity score of u is in the range of [0, 1] as well.

To properly integrate the proximity score with the discriminating strength of u with respect to c, PDFI computes a ranking score based on the strength rank of u among all candidate features of c (i.e. *RankScore*, ref., Step 6)[1]. The final strength of u is simply the sum of the ranking score and the proximity score (ref., Step 7). The strength estimation is based on the observation that, a good diagnosis factor u for a disease c should be able to (1) discriminate c from other diseases *and* (2) individually appear near to other candidate features for c. Therefore, the combination of the strength rank and the proximity score of a candidate feature u may produce a strength to indicate the possibility of u serving as a diagnosis factor for disease c. Based on the strengths estimated for candidate features, PDFI may rank good diagnosis factors higher.

[1] PDFI does not directly employ the discriminating strength of u, since different discriminative factors identifiers may produce different scales of strengths, making their integration with the proximity score more difficult.

3 Empirical Evaluation

An experiment is conducted to empirically evaluate the contribution of PDFI. The experiment aims at measuring how PDFI may enhance a state-of-the-art discriminative factors identifier in ranking candidate factors using a real-world database of medical texts for a broad range of diseases. Table 2 summarizes the main experimental setup, which is to be described in the following subsections.

Table 2. Experimental setup to evaluate PDFI

Item	Setting
(1) Dictionary of medical terms	Medical terms is collected from the 2011 MeSH, with each MeSH term, its retrieval equivalence terms (terms in its "entry" field) are included as well, resulting in a dictionary of 164,354 medical terms.
(2) Source of experimental data	(A) Experimental texts are collected from MedlinePlus by checking all the diseases for which MedlinePlus tags diagnosis/symptoms texts, resulting in a text database of 420 medical texts for 131 diseases. (B) Each medical text is manually read and cross-checked to extract target diagnosis factor terms from the texts, resulting in 2,797 target terms (that are included in MeSH as well) from the 420 medical texts for 131 diseases.
(3) Underlying discriminative factors identifiers	The chi-square (χ^2) feature scoring technique is employed as the underlying factor identifier, which produces a discriminating strength for each feature (candidate factor) with respect to each disease, and for each disease, all positively-correlated features are sent to PDFI for re-ranking.
(4) Evaluation criterion	Mean Average Precision (MAP) is employed as the criterion, which measures how target diagnosis factors are ranked high for the medical expert to check and validate.

3.1 Experimental Data and Resources

Recognition of biomedical terms is a basic step of diagnosis factor identification, since a factor is actually expressed with a valid biomedical term. It is also known as the problem of named entity recognition, and several previous studies focused on the recognition of new terms not included in a dictionary (e.g., [12][16]). In the experiment, since we do not aim at dealing with new terms, we employ a dictionary of medical terms to recognize medical terms, which are then treated as candidate factors. All terms in a popular and up-to-date database of medical terms MeSH (medical subject headings) are employed.[2] By including all terms and their retrieval equivalence terms,[3] we have a dictionary of 164,354 medical terms, which should be able to cover

[2] We employ 2011 MeSH as the dictionary, which is available at `http://www.nlm.nih.gov/mesh/filelist.html`

[3] Retrieval equivalence terms of a term are from the field of "entry" for the term, since they are equivalent to the term for the purposes of indexing and retrieval, although they are not always strictly synonymous with the term (ref. `http://www.nlm.nih.gov/mesh/intro_entry.html`)

most diagnosis factor terms. A term may be a candidate factor term only if it is listed in the dictionary, making the systems able to focus on valid medical terms.[4]

On the other hand, to test the factor identification techniques, we need experimental medical texts from which the techniques identify diagnosis factors for individual diseases. Since there is no benchmark database for the purpose, we collect the medical texts from MedlinePlus,[5] which aims at providing reliable and up-to-date information for diseases. We examine all diseases listed in MedlinePlus, and download the texts that MedlinePlus tags as diagnosis/symptoms texts for the diseases.[6] Each text is for a disease. It is manually read and cross-checked to extract the passages that talk about the target diagnosis factors for the corresponding disease. The dictionary noted above is then used to filter out those terms not in MeSH. We then remove those texts that do not have target diagnosis factors terms and those diseases that do not have any texts with target diagnosis factors terms. There are thus 420 medical texts for 131 diseases,[7] and 2,797 target diagnosis factor terms for the diseases. The texts may thus be a suitable data to comprehensively evaluate PDFI in a broad range of diseases.

A few steps are conducted to preprocess the medical terms, including transforming all characters into lower case, replacing non-alphanumeric characters and "'s" with a space character, and removing stop words.[8]

3.2 The Underlying Discriminative Factors Identifier

We implement χ^2 (chi-square) as the underlying factor scoring and identification technique. It is a state-of-the-art technique that was routinely employed (e.g., [4][14]) and shown to be one of the best feature (factor) scoring techniques [14]. For a term t and a category (disease) c, $\chi^2(t,c) = [N \times (A \times D - B \times C)^2] / [(A+B) \times (A+C) \times (B+D) \times (C+D)]$, where N is the total number of documents, A is the number of documents that are in c and contain t, B is the number of documents that are not in c but contain t, C is the number of documents that are in c but do not contain t, and D is the number of documents that are not in c and do not contain t. Therefore, $\chi^2(t,c)$ indicates the strength of correlation between t and c. We say that t is *positively correlated* to c if $A \times D > B \times C$; otherwise t is *negatively correlated* to c. A term may be treated as a

[4] Actually, MeSH provides a tree (http://www.nlm.nih.gov/mesh/trees.html) containing various semantic types that are related to diagnosis factors, including "Phenomena and Processes" (tree node G), "diagnosis" (tree node E01), and "disease" (tree node C). However, diagnosis factor terms may still be scattered around the MeSH tree. For example, the symptom term "delusions" is under the MeSH category of Behavior and Behavior Mechanisms (tree node F01) → Behavior → Behavioral Symptoms → Delusions. Therefore, we employ the whole list of MeSH terms as the dictionary.

[5] Available at `http://medlineplus.gov`

[6] The texts are collected by following the sequence: MedlinePlus → "Health Topics" → Items in "Disorders and Conditions" → Specific diseases → Texts in "diagnosis/symptoms." We skip those diseases that have no "diagnosis/symptoms" texts tagged by MedlinePlus.

[7] The 131 diseases fall into 7 types of disorders and conditions: cancers; inflections; mental health and behaviors; metabolic problems; poisoning, toxicology, environmental health; pregnancy and reproduction; and substance abuse problems.

[8] A stop word list by PubMed is employed:
`http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp&rendertype=table&id=pubmedhelp.T43`

candidate factor term for c only if it is positively correlated to c. By comparing the performances of the systems *before* and *after* PDFI is applied (named χ^2 and χ^2+PDFI, respectively), we may measure the contribution of PDFI to χ^2.

3.3 Evaluation Criteria

Both the underlying discriminative factor identifier (i.e., χ^2) and the enhanced version of the identifier (i.e., χ^2+PDFI) produce strengths for candidate factor terms, and hence produce two lists of ranked terms. Therefore, by measuring how target diagnosis factor terms appear in the two ranked lists, we may evaluate the quality of the strength estimations by χ^2 and χ^2+PDFI, and accordingly measure the contribution of PDFI to the underlying discriminative factor identifier. Therefore, *Mean Average Precision* (MAP) is employed as the evaluation criterion:

$$MAP = \frac{\sum_{i=1}^{|C|} P(i)}{|C|}, \quad P(i) = \frac{\sum_{j=1}^{m} \frac{j}{f_i(j)}}{k} \tag{1}$$

where $|C|$ is the number of diseases (i.e., 131), k is number of target diagnosis factor terms for the i^{th} disease, and $f_i(j)$ is the number of candidate terms whose ranks are higher than or equal to that of the j^{th} target term for the i^{th} query. That is, $P(i)$ is actually the *average precision* (AP) for the i^{th} disease, and MAP is simply the average of the AP values for all diseases.

Note that when computing $P(i)$ for a disease, we employ the number of target terms (i.e., k) as the denominator, making it able to consider the percentage of target terms in the ranked list (i.e., a sense similar to *recall*). Therefore, if the i^{th} disease has many target terms but only very few of them are in the ranked list, $P(i)$ will be quite low no matter how the few terms are ranked in the list.

3.4 Results

MAP values of χ^2 and χ^2+PDFI are 0.2136 and 0.2996. After conducting a two-tailed and paired t-test on the average precisions on the 131 individual diseases (ref., P(i) in Equation I), we find that the performance difference is statistically significant with 99% confidence level, indicating that PDFI successfully enhances χ^2 in ranking target diagnosis factors higher. Given that χ^2 is a state-of-the-art factor scoring technique, the enhancement provided by PDFI is of both practical and technical significance. If there exists another scoring technique that may produce better ranking than χ^2, PDFI may collaborate with it as well.

To illustrate the contribution of PDFI, consider the disease 'parasitic diseases' that is of the disorder type of infections, and there are two texts for it. Average precision for the disease achieved by χ^2 and χ^2+PDFI are 0.3003 and 0.3448, respectively. When compared with χ^2, χ^2+PDFI promotes the ranks of several target diagnosis factors such as 'parasite,' 'antigen,' 'diarrhea,' and 'MRI scan,' since they appear at some place(s) where more other candidate terms occur in a nearby area. On the other hand, χ^2+PDFI also lowers the ranks of a few target diagnosis factors such as

'serology.' A detained analysis shows that 'serology' only appears at one place where the author of the text used lots of words to explain 'serology' with very few candidate terms appearing in the nearby area. As anther example, consider the disease 'cervical cancer' that is of the disorder type of cancers, and there are two texts for it. Average precision for the disease achieved by χ^2 and χ^2+PDFI are 0.1222 and 0.4071, respectively. χ^2+PDFI promotes the ranks of several target factors such as 'colposcopy' and 'vagina.' Overall term proximity contexts are helpful in promoting the ranks of target diagnosis factors.

4 Conclusion and Future Work

Diagnosis factors to discriminate a broad range of diseases are the fundamental basis on which intelligent systems may be built for diagnosis decision support, diagnosis skill training, medical research, and health education. A diagnosis knowledge map that consists of the relationships between diseases and diagnosis factors is an ontology for the systems. Its construction requires the support provided by a text mining technique since (1) the diagnosis factors should be discriminative among the diseases, (2) the diagnosis factors may evolve as medical findings evolve, and (3) medical findings are often recorded in texts. We thus develop a technique PDFI that employs term proximity contexts to significantly enhance existing factor identification techniques so that they may be more capable of extracting diagnosis factors from a given set of medical texts. The contribution of PDFI is based on the observation that authors of a medical text for a disease tend to describe the diagnosis factors of the disease (if any) in a nearby area of the text. Based on the research results, we are exploring several interesting issues, including automatic collection of medical texts for diagnosis factor extraction, computer-supported validation of diagnosis factors, and visualization of the diagnosis knowledge map for interactive map exploration. A complete, reliable, and evolvable diagnosis knowledge map may be expected to be shared among users and systems in various medical applications.

Acknowledgments. This research was supported by the National Science Council of the Republic of China under the grants NSC 99-2511-S-320-002.

References

1. Cummins, R., O'riordan, C.: Learning in a Pairwise Term-Term Proximity Framework for Information Retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, USA, pp. 251–258 (2009)
2. Domedel-Puig, N., Wernisch, L.: Applying GIFT, a Gene Interactions Finder in Text, to Fly Literature. Bioinformatics 21, 3582–3583 (2005)
3. Gerani, S., Carman, M.J., Crestani, F.: Proximity-Based Opinion Retrieval. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, pp. 403–410 (2010)

4. Himmel, W., Reincke, U., Michelmann, H.W.: Text Mining and Natural Lan-guage Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums. Journal of Medical Internet Research 1(3), e25 (2009)
5. Kim, S., Yoon, J., Yang, J.: Kernel Approaches for Genic Interaction Extraction. Bioinformatics 24, 118–126 (2008)
6. Liu, R.-L.: Text Classification for Healthcare Information Support. In: Proceedings of the 20th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, pp. 44–53. Kyoto University, Kyoto (2007)
7. Mladeniá, D., Brank, J., Grobelnik, M., Milic-Frayling, N.: Feature Selection Using Linear Classifier Weights: Interaction with Classification Models. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 234–241 (2004)
8. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature. Bioinformatics 17, 155–161 (2001)
9. Özgür, A., Vu, T., Erkan, G., Radev, D.R.: Identifying Gene-Disease Associations Using Centrality on a Literature Mined Gene-Interaction Network. Bioinformatics 24, i277–i285 (2008)
10. Suebnukarn, S., Haddawy, P.: Modeling individual and collaborative problem-solving in medical problem-based learning. User Modeling and User-Adapted Interaction 16, 211–248 (2006)
11. Svore, K.M., Kanani, P.H., Khan, N.: How Good is a Span of Terms? Exploiting Proximity to Improve Web Retrieval. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, pp. 154–161 (2010)
12. Takeuchi, K., Collier, N.: Bio-medical Entity Extraction Using Support Vector Machines. Artificial Intelligence in Medicine 33, 125–137 (2005)
13. Temkin, J.M., Gilder, M.R.: Extraction of Protein Interaction Information from Unstructured Text Using a Context-Free Grammar. Bioinformatics 19, 2046–2053 (2003)
14. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee, pp. 412–420 (1997)
15. Zhao, J., Yun, Y.: A Proximity Language Model for Information Retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, USA, pp. 291–298 (2009)
16. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C.: Recognizing Names in Biomedical Texts: a Machine Learning Approach. Bioinformatics 20, 1178–1190 (2004)

A Method for User Profile Adaptation in Document Retrieval

Bernadetta Mianowska and Ngoc Thanh Nguyen

Wroclaw University of Technology,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
Bernadetta.Mianowska@pwr.wroc.pl
Ngoc-Thanh.Nguyen@pwr.edu.pl

Abstract. On the Internet the number of Web pages and other documents has grown so fast that it is very hard to find needed information. Search engines are still improving their retrieval methods but still many irrelevant documents are presented in the results. A solution to this problem is to get to know the user, his interests, preferences and habits and use this information in retrieval process. In this paper a user profile and its adaptation method is proposed. To evaluate the proposed method, simulation of user behaviour is described. Performed experimental evaluation shows that the distance between created user profile and user preferences is decreasing with subsequent actualization processes steps.

Keywords: user profile adaptation, simulation of user behaviour, document retrieval, personalization.

1 Introduction

Nowadays, with rapid growth of the Internet, large amount of information is available. Each user can use this data but the problem arises because of its overload. There is so much data and information in the Internet, that normal user is not able to find needed information in a reasonable time. Many search engines are still improving their methods of information retrieval but the user would like to get only relevant information in short time. The problem may come not only from the search engine deficiencies (such information may not exist) but also from user's lack of knowledge or skills. User may not know how to formulate a query, i.e. when he does not know about some aspects of problem or he is a new to the system [1].

Usually user enters very few words to the search engine and expects that the system will know real users' needs. To meet such expectations many systems are trying to get more information about the user and use that information during the searching process. The most popular way is to ask the user about his interests, preferences, hobbies, etc. and save them in the user profile. More complex methods are based on observation of user's activity in the system and adapt according to them. Information gathered in user profile can be used to establish real context of user's query or disambiguate the sense of it.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 181–192, 2011.

Either explicit or implicit methods have advantages and disadvantages. Explicit methods require user's time to fill some questionnaire or look through a large set of documents to explore user's interests' areas. On the other hand, implicit method does not engage the user but need more time to decide whether he is interested in some field or not [14]. In both situations it is possible that system would not gather information about real users' interests and preferences.

The most important problem in user profiling is the fact that users' interests, preferences and habits are changing with time. To keep user profile up-to-date it is necessary to use adaptation method based on user feedback [7], e.g. documents that user opens, saves or print. The user judges relevance of documents and it has strong influence to profile adaptation.

In this paper we propose a user profile and describe an adaptation method. Proposed solutions are evaluated using simulated user. The main advantage of not using real people to the experiment is to save time but on the other hand, it requires many intuitive assumptions about user's behaviour in searching situation.

This paper is organized as follows: Section 2 contains an overview of personalization methods, problems with keeping profile up-to-date and adaptation techniques. Section 3 proposes user profile and adaptation method. In Section 4 we describe how user behaviour in searching process is simulated. Section 5 presents our experiment and discussion about the first results. Conclusions and future works are contained in Section 6.

2 Related Works

Building the user profile in personalization systems is not a new idea and many applications in this area are still developed to recommend better documents to the user. Many search engines such as Yahoo, Google, MSN, and AltaVista, are developed to meet users' needs, but they do not satisfy the various users' needs in real world. [6]. The main problem is that getting to know the user: his interests, preferences, habits, etc. and saving this knowledge in a user profile is not sufficient. Information about the user can become out-of-date because he has changed his preferences or is getting to know a new discipline.

Despite efficient information retrieval technologies, users are still not satisfied with the search process and the results presented. Studies have shown that when user is not familiar with the topic, he enters very few query terms. As a result, user is often inundated with a large amount of links returned due to the generality of the terms used during a search. It can be alleviated by providing more user information when the search was performed. [4]. System with user profile can help user in searching process by extending user's query or by disambiguating the context of it.

Approaches to user profiling can be divided into two groups according to the way of profile generation: static and dynamic. Static profile approach means that users' preferences, interests or other values are static after created the user profile. Most

portal systems use the static profile approach to provide personalized information. In that case user profile can have incorrect information about the user because users' preferences have changed. To address this problem, dynamic profiles are created and various learning techniques, such as Bayesian classifiers, neural networks, and genetic algorithms have been utilized for revising user profiles in several research studies resulting in various levels of improvement [6]. A cognitive user model presented in [11] is used to characterize optimal behavioral strategies for information search using a search-engine.

In many systems, authors assume the existence of long-term and short-term interests [10]. The first category of terms are more constant and theirs changes are rarely. Short-term can change very quickly and usually are not so important for the user. In this context, the aim of personalization process is to differentiate those terms and save information that is still relevant to users' real information needs and change the influence of this information during searching process.

To build useful user profiles, many acquisition methods are generated. User enters a query and looks through the results. The system can ask the user how much each document was relevant to the query (explicit method) or observe user actions in the system (implicit method). Authors of [13] list a number of indicator types that can show if user is interested in document or not. There are following indicators: explicit – user selects from scale, marking – bookmark, save, print; manipulation – cut/paste, scroll, search; navigation – follow link, read page; external – eye movement, heart rate; repetition – repeated visits; negative – not following a link. All those features, except the first one, can be used in an implicit manner.

Different systems are generating different user profile structures and it has direct influence on adaptation and modification methods. The user profile can save information about the user in the following forms: list of historical activity, vector space model (Boolean model or with weighted terms), weighted association network, user-document matrix, hierarchical structures (tree) or ontologies [9]. Actualization of user profile depends on the profile structure. In most systems user interests, preferences, etc. are saved as set of terms with appropriate weights and terms can be linked by some relationships. Adaptation method means changing values of those weights by, sometimes complicated, mathematical formula, e.g. statistical regression or Bayesian model [14]. The way of weights modification is as important as the data, based on which, modification is done.

Important issue in profile adaptation process is not only to add new information about user but also to judge if data existing in profile is valid. The first systems like Sift NetNews obligated users to manually modify their profile by adding or removing some interests [10]. Newer solutions are based on relevance feedback and user observation. When user is not interested in some area for a long time, this area has lower influence with time and is getting unimportant in user profile. Biologically, the reasons for the forgetting model are described in [4]: amount of unrepeated information remembered by a person is exponentially decreasing with time.

User profile modification is a difficult task because information gathered about the user can not be reliable in the sense of real user information needs. That is the reason

why not only positive feedback is taken into account but also the negative one [12]. The user has selected only a few documents from usually large list, and the other documents are omitted because of many possible reasons: user has found needed information in previous document, user is not interested in the other documents because they are irrelevant, user has not found it and is trying to reformulate a query or simply user does not have time to open the remaining documents.

3 User Profile and Adaptation Method

In this section we present our model for the user profile and a method of the adaptation.

Agent-based Personal Assistant is a system consisting of three modules: Personalization, Metasearch and Recommendation. The Personalization component gathers information about users' activity, interests and relevant documents. Meta Search part is responsible for finding the best search engines and sending there queries, collecting answers from all the used sources and transferring them to Recommendation part. The task of the Recommendation component is used to select and sort retrieved documents and to present them to the user [8], [9].

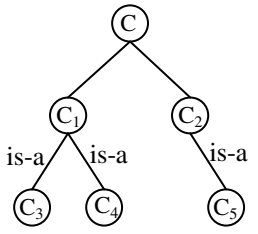
The main aim of Personalization module is to get to know the user and to build an appropriate user profile based on collected data. To guarantee that the user profile can be effectively used in the retrieval process, it should stay up-to-date. Personalization module is observing users' activities in the system and modifying the user profile.

As presented in our previous work [9] the user profile can be presented as a tree structure – acyclic and coherent graph $G = (V, E)$, where V is a set of nodes containing the vector of weighted terms describing concepts and a time stamps when the user was interested in those concepts for the last time. The E is a set of edges representing "is-a" relation between two nodes taken from WordNet ontology [17]. If two nodes v_1 and v_2 are connected by edges $e(v_1,v_2)$, it means that concept from node v_2 is a kind of concept in node v_1. Terms contained in the same concepts are synonyms. Figure 1 presents an example of user profile structure.

The root of this tree is not a concept in WordNet meaning but only an artificial node. The nodes of the first level (roots' children) are the main areas of user interests or concepts connected with them. The profile of the user with many different interests will be broad (many children of the root) and the deeper the user profile is, the more specialized information about the user the system have.

The terms in the user profile are obtained from users' queries. The intuition in this situation is that if the user is asking about some information, he is interested in it and even if there are other terms connected with users' ones, it is not obvious that the user is interested also in those additional terms.

To perform dynamic adaptation process the system observes the user and saves his queries and documents that were received as results in each session. Session is a set of users' action in the system from the opening of the system to its closure. After the



C_i=<c_i, t_i>

c_i – *set of weighted synonyms describing i-th concept:*

$$c_i = \{(k_1^{(i)}, w_1^{(i)}), (k_2^{(i)}, w_2^{(i)}), \ldots, (k_{n_i}^{(i)}, w_{n_i}^{(i)})\}$$

$k_j^{(i)}$ – *j-th synonym*

$w_j^{(i)}$ – *weight of j-th synonym*

t_i – *time stamp of last appearance of concept* c_i *or its subconcepts in user's query*

Fig. 1. A structure of user profile

session, from the set of documents that user has received and considered as relevant to appropriated query, the mean value of each terms from users' queries is calculated:

- Documents set from session with appropriate query:

$$D(s) = \{(q_i, d_{i_j}^{(s)})\}. \tag{1}$$

where: s is session's number, i is query's number and i_j is a number of documents relevant to query q_i;

- Weight of k_l-th term after s session:

$$w_d^{(s)}(k_l) = \frac{1}{n_s} \sum_{i=1}^{n_s} w_{d_i}^{(s)}(k_l). \tag{2}$$

where: $w_d^{(s)}(k_l)$ is average value of weights in documents set and $w_{d_i}^{(s)}(k_l)$ is weight of term in single document in current session.

From each session system obtains a set of terms that were used in queries in this session. To calculate the weights of those terms in relevant documents, some measure of term importance in the document should be used. The most popular method to extract index terms from the document is statistical term weighting measure frequency-inverse document frequency (TF-IDF). In research collections, documents are often (but not always) described by some keywords added by the authors. Using existing keywords is more adequate than using terms artificially extracted from the document because the authors know the best way to describe their work. Some danger can be hidden in that approach when authors add words or terms that are not present in dictionary or ontology used by the system (e.g. added word is a proper name or too specialized in a very narrow research area).

After the session, new set of weighted terms are calculated. To update the weights of existing terms in profile, the change of user interests can be calculated as:

- Relative change of user interests in term k_l after the session s:

$$\Delta w_{k_l}(s) = \begin{cases} \dfrac{w_{d_i}^{(s)}(k_l) - w_{d_i}^{(s-1)}(k_l)}{w_{d_i}^{(s-1)}(k_l)}, & if\ w_{d_i}^{(s-1)}(k_l) > 0 \\ w_{d_i}^{(s)}(k_l), otherwise \end{cases}. \quad (3)$$

where: $w_{d_i}^{(s)}(k_l)$ is a weight of term k_l after current session and $w_{d_i}^{(s-1)}(k_l)$ is a weight of term k_l after previous session.

User adaptation process is a function of two arguments: current weight of term in user profile $w_{k_l}(s)$ and relative change of user interests in two last sessions $\Delta w_{k_l}(s)$:

$$w_{k_l}(s+1) = f(w_{k_l}(s), \Delta w_{k_l}(s)). \quad (4)$$

The following formula is proposed to calculate the weight of term k_l after the session s:

$$w_{k_l}(s+1) = \alpha \cdot w_{k_l}(s) + (1-\alpha) \cdot \left(\frac{A}{1 + B \cdot exp(-\Delta w_{k_l}(s) + C)} \right). \quad (5)$$

where: $w_{k_l}(s+1)$ is a weight of term k_l in user profile in session $s+1$; A, B, C and α are parameters that should be attuned in experimental evaluation.

When the user is not more interested in some terms, theirs weights are decreasing and if this weight is smaller than some assumed threshold, such unimportant term is deleted from user profile.

In the next sections, we show that proposed adaptation method is effective in terms of the Euclidean measure. After the 3rd, 6th and 9th block of sessions, user profile will be updated and the distance between created user profile and user preferences will be calculated in those steps. The distance in subsequent stages will decrease so the created user profile will become closer to user preferences.

4 User Simulation

To perform an experiment instead of using a real user we would like to simulate user behaviour. We assume that users' preferences are described by set of 10 terms randomly selected from a thesaurus T. To each term a normalized weight w_i is generated $w_i \in (0{,}5; 1]$, $i \in \{1,2,...,10\}$.

Clarkea et al. [3] checked that in many retrieval systems more than 85% of the queries consisted of three terms or less. In our experiment user queries will be generated by selecting 3 terms from user preferences and will be sent to the information retrieval system. It is possible that system will not return any document to

entered query because selected terms can be inconsistent [5] or simply such document can not exist in database.

User preferences can change with time so after each 10 sessions one term with the lowest weight will be deleted and another term from thesaurus T will be chosen and added to user preferences with random weight. The weights of other terms in users' preferences set will be changed by at most 10% [2], but still they will fulfill the condition $w_i \in (0,5;1],\ i \in \{1,2,...,10\}$. User preferences can be presented in the following form (with weights in increasing order):

$$\begin{gathered} Pref = ((t_1,w_1),(t_2,\ w_2),\ ...,(t_{10},\ w_{10})) \\ 0,5 \le w_1 \le w_2 \le ... \le w_{10} \le 1 \end{gathered}. \tag{6}$$

The most important feature in real user behaviour is to judge whether retrieved document is relevant to entered query or not. To do it automatically, additional weights are needed to each term of user query q and document keywords d. The best indexes are added to the document by its authors: $d = (i_1, i_2, ..., i_m)$; where m depends on how many indexes author(s) entered.

In our experiment relevance function will be used. Relevance function is calculated as a sum of differences between positions of term in the user query and position of the same term or its synonym in document indexes divided by maximal sum of those differences. The numerator of this value of proposed relevance measure can be treated as the number of operations needed to set those terms in the same order as they are in user query and on the beginning of indexes list:

$$dist(q,d) = \frac{\sum_i |pos(term_i^q) - pos(term_i^d)|}{\sum_j |num\_of\_pos(j) - 1|}. \tag{7}$$

where: $pos(term_i^q)$ is a position of i-th term in user query; $pos(term_i^d)$ is a position of *i-th* term in documents' indexes and $num\_of\_pos(j)$ is a number of indexes in *j-th* document.

5 Experimental Evaluation

In this section, we discuss our methodology to perform the experiments to evaluate the effectiveness of personalization method. The aim of the experiment is to show that the distance between user preferences and generated profile is getting smaller when actualization method is applied.

The experiment was performed according the following plan:

1. Determine a domain and set of terms T that will be used in experiment. Determine user preferences *Pref*.

2. Generate a query q – select randomly 3 terms from user preferences set $\{t_1, t_2, ...,t_{10}\}$, set these terms in proper order according to decreasing weights and send query to search engine.
3. From the return list get keywords $d = (i_1, i_2, ...,i_m)$; m is a number of indexes.
4. Calculate distances $dist(q,d)$ between each retrieved document and user query using formula (7).
5. Mark documents that meet the condition $dist(q,d) \geq \rho;\ \rho \in (0;1)$ as 'relevant'; $D = \{d_1^{(s)}, d_2^{(s)},..., d_{n_s}^{(s)}\}$
6. After 3 blocks of 10 session update the profile using adaptation method presented in equation (5).
7. Calculate the distance between user profile and user preferences using Euclidean distance between those two extended vectors.

Experiment was performed on ACM Digital Library [15].

The following assumptions were made for the experiment: we assume the values of weights of documents' indexes are known. The first keyword is the most important and has weight $w(i_1)=1$, weight of the next keyword is $w(i_2)=0,9$ and so on. The sixth and next keywords will have weights $w(i_m)=0,5$:

$$\begin{aligned} w(i_1) &= 1 \\ w(i_2) &= 0{,}9 \\ w(i_3) &= 0{,}8 \\ &\vdots \\ w(i_m) &= 0{,}5 \ for\ m \geq 6 \end{aligned} \quad . \tag{8}$$

The terms from user query do not need weights. User query contains three terms and we assume that each of them is equally important. The order of those terms in query is taken into account by used search engine.

All methods described in previous sections were implemented in Java environment (J2SE). The set of terms T was obtained from the ACM Computing Classification System [16] and to user preferences *Pref* the following terms were selected with randomly weights:

Pref(*0*) = {(*database, 0.81*), (*network, 0.65*), (*knowledge, 0.63*), (*security, 0.9*), (*management, 0.86*), (*learning, 0.83*), (*hybrid, 0.86*), (*theory, 0.71*), (*simulation, 0.5*), (*semantic, 0.53*)}

From the set *Pref*(*0*) users' queries were generated by random selection of three out of 10 terms. In each session we assume 5 users' queries. After each block of 10 sessions, the term with the lowest weight was replaced by new term from T and new weight for its term was generated. The sample set of users' queries with the results (indexes of

retrieved documents) obtained from the ACM Portal from one session is presented in Table 1. Each document was checked by the relevance function and from the set of relevant documents after 3 blocks of sessions, the first user profile was generated according to procedure described in Section 3.

After the very first trials, the parameters were tuned and set as follows: $\rho = 0.3$; $A = 1.0$, $B = 1.0$ and $C = 3.0$. To the experiment 9 blocks of sessions was prepared. After first 3

Table 1. A sample of user session: queries and results

User query	Indexes of documents obtained for query
theory knowledge learning	knowledge management, media theory, multimedia process model, technology enhanced learning
	knowledge, knowledge management, learning organisation, management, organisational learning, theory
	collaboration, corporate universities, customer relations management (CRM), disciplinarity, educational theory, knowledge management, multimedia, organizational learning, research, training, usability studies, user-centered design
	database courseware, knowledge and skills, multimedia, tool-mediated independent learning, virtual apprenticeship theory
knowledge network security	OSPF attacks, event correlation, knowledge-based IDS, link-state routing protocol security, real-time misuse intrusion detection, real-time network protocol analysis, timed finite state machine
	computer-network security, knowledge acquisition, knowledge-based temporal abstraction, malicious software, temporal patterns
	AI reasoning, Vijjana model, cloud computing, interactive knowledge network, security, semantic web
	base stations, entity authentication, guillou-quisquater protocol, security protocols, sensor and ad hoc networks, wireless security, zero-knowledge protocol
	IHMC, MAST, concept maps, knowledge models, mobile agents, network security
learning database theory	learning theory, non-interactive database privacy
	database courseware, knowledge and skills, multimedia, tool-mediated independent learning, virtual apprenticeship theory
hybrid simulation management	failure management, hybrid embedded software, model-in-the-loop, quality assurance, simulation, simulink, testing
database management learning	LMS, Oracle, PHP, authentication, automated, database, email, instructor, lab, learning management system, online, registration, reservation, roster, training, workshop
	Case-based learning, UML, conceptual modeling, database management systems, design, education, informatics, information retrieval, lucene, postgresql

Table 2. User profile generated and updated in subsequence of 3 blocks

User profile after 3rd block		**User profile after 6th block**		**User profile after 9th block**	
term	*weight*	*term*	*weight*	*term*	*weight*
network	0,13387	network	0,28212	network	0,00000
simulation	0,14519	simulation	0,00000	simulation	0,00000
theory	0,13741	theory	0,28860	theory	0,00000
hybrid	0,12670	hybrid	0,27633	hybrid	0,41713
semantic	0,14311	semantic	0,00000	semantic	0,00000
knowledge	0,13035	knowledge	0,28021	knowledge	0,41639
learning	0,12823	learning	0,27317	learning	0,41853
security	0,13483	security	0,27616	security	0,41331
database	0,12458	database	0,26940	database	0,00000
management	0,13163	management	0,28059	management	0,41704
dynamic	0,12401	dynamic	0,27281	dynamic	0,41474
collaborative	0,11686	collaborative	0,00000	collaborative	0,00000
		storage	0,18713	storage	0,34458
		intelligence	0,18899	intelligence	0,34744
		distributed	0,18845	distributed	0,34557
				agent	0,18808
				cognitive	0,18727
				modeling	0,18675

blocks the user profile was generated and after 6-th and 9-th block, this profile was updated. Table 2 presents the user profile obtained after each update in the vector form.

To compare created and updated user profile with user preferences, the Euclidean measure was used. In the subsequent blocks of session, new terms can be added to the user profile so we need to modify this measure. To compare two vectors with different numbers of dimentions, Euclidean measure was normalized by dividing calculated value by the maximal length of vector with normalized coordinates. The results are presented in Fig.2. The distance between user preferences and created user profile is decreasing with subsequent updates.

Trends in Euclidead measure show that adapting user profile using proposed algorithm is effective and created and updated profile is becoming more similar to preferences vector, in spite of the fact that preferences are slightly changed with time.

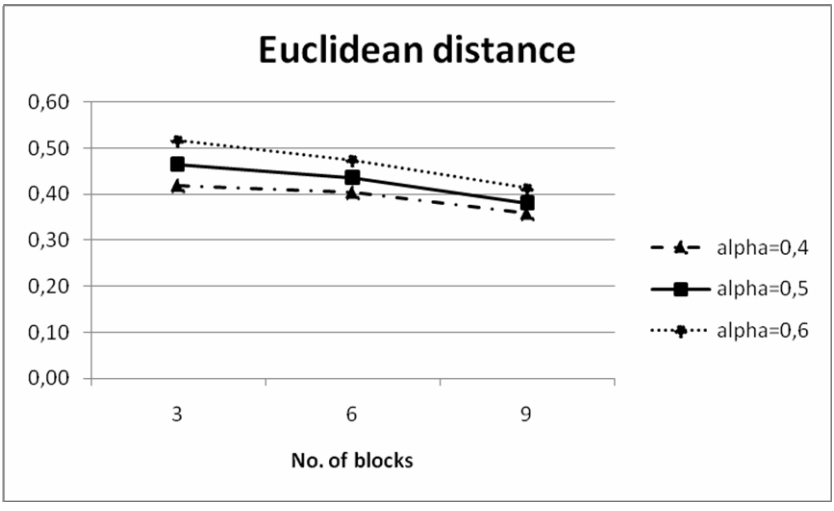


Fig. 2. Euclidean distance between user preferences and the created user profile

6 Conclusions

In this paper we present a proposition of user profile and method of its adaptation. To avoid time and work-consuming experiments with real users, the way of user simulation was presented. Performed experiment shows that updating the profile according to proposed formula is an effective way of user profiling and the adapted profile becomes more similar to user preferences.

Acknowledgments. This research was partially supported by European Union within European Social Fund under the fellowship no. MK/SN/160/III/2010/U and by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

References

1. Aldous, K.J.: A System for the Automatic Retrieval of Information from a Specialist Database. Information Processing & Management 32(2), 139–154 (1996)
2. Carreira, R., Crato, J.M., Gonçalves, D., Jorge, J.A.: Evaluating Adaptive User Profiles for News Classification. In: IUI 2004, Portugal (2004)
3. Clarkea, C.L.A., Cormackb, G., Tudhope, E.A.: Relevance ranking for one to three term queries. Information Processing & Management 36, 291–311 (2000)
4. Dingming, W., Dongyan, Z., Xue, Z.: An Adaptive User Profile Based on Memory Model. In: The Ninth International Conference on Web-Age Information Management. IEEE, Los Alamitos (2008)
5. Iivonen, M.: Consistency in the Selection of Search Concepts and Search Terms. Information Processing & Management 31(2), 173–190 (1995)

6. Jeon, H., Kim, T., Choi, J.: Adaptive User Profiling for Personalized Information Retrieval. In: Third 2008 International Conference on Convergence and Hybrid Information Technology. IEEE, Los Alamitos (2008)
7. Kobsa, A.: User Modeling and User-Adapted Interaction. In: Conference Coempanion CID 1994 (1994)
8. Maleszka, M., Mianowska, B., Nguyen, N.T.: Agent Technology for Information Retrieval in Internet. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. LNCS, vol. 5559, pp. 151–162. Springer, Heidelberg (2009)
9. Mianowska, B., Nguyen, N.T.: A Framework of an Agent-Based Personal Assistant for Internet Users. In: Jędrzejowicz, P., Nguyen, N.T., Howlet, R.J., Jain, L.C. (eds.) KES-AMSTA 2010. LNCS (LNAI), vol. 6070, pp. 163–172. Springer, Heidelberg (2010)
10. Mitra, M., Chaudhuri, B.B.: Information Retrieval from Documents: A Survey. Information Retrieval 2, 141–163 (2000)
11. O'Brien, M., Keane, M.T.: Modeling User Behavior Using a Search-Engine. In: IUI 2007, USA (2007)
12. Pan, J., Zhang, B., Wang, S., Wu, G., Wei, D.: Ontology Based User Profiling in Personalized Information Service Agent. In: 17th International Conference on Computer and Information Technology (2007)
13. Shen, X., Tan, B., Zhai, C.X.: Implicit User Modeling for Personalized Search. In: CIKM 2005, Germany (2005)
14. Story, R.E.: An Explanation of the Effectiveness of Latent Semantic Indexing by Means of a Bayesian Regression Model. Information Processing & Management 32(3), 329–344 (1996)
15. ACM Digital Library, `http://portal.acm.org/dl.cfm`
16. ACM Classification, `http://www.acm.org/about/class/ccs98`
17. WordNet Ontology, `http://wordnet.princeton.edu/`

Intelligent Image Content Description and Analysis for 3D Visualizations of Coronary Vessels

Mirosław Trzupek, Marek R. Ogiela, and Ryszard Tadeusiewicz

AGH University of Science and Technology,
Institute of Automatics, Laboratory of Biocybernetics,
30 Mickiewicza Ave, 30-059 Krakow, Poland
{mtrzupek,mogiela,rtad}@agh.edu.pl

Abstract. The paper will discuss in detail the new possibilities for making linguistic description and semantic interpretations of 64-slice spiral CT coronary vessels visualizations with the use of AI linguistic formalisms and especially ETPL(k) graph grammar. Current research shows that a significant part of diagnostic imaging, including of coronary arteries, is still difficult to automatically assess using computer analysis techniques aimed at extracting information having semantic meaning. The proposed syntactic semantic description makes it possible to intelligently model the examined structure and then to automatically find the locations of significant stenoses in coronary arteries and identify their morphometric diagnostic parameters.

Keywords: image semantic description, syntactic pattern analysis, medical image understanding, Computer-Aided Diagnosis (CAD).

1 Introduction

Coronary heart disease are among the most important public health problems in developed countries. It is one of the major cause of death in the world [21]. It is desirable to search for tools capable of helping the physician take therapeutic decisions (second-opinion assistance). What is more, the impressive progress in apparatuses for acquiring medical images means that new generations of diagnostic apparatuses enter the market. As a result, the images analyzed so far, familiar to the physician, are replaced with other, better ones. Although the latter contain more information, they are unfortunately unfamiliar. In the present situation it is also increasingly difficult to be an expert who could efficiently assess all the currently available types of medical images. The reason is that the wide range of diagnostic apparatuses for visualizing the same organ makes it possible to present the organ in various visual forms depending on the technique used to acquire the image. All these and many other circumstances have led the authors of this publication to carry out wide-ranging research to find new solutions that could help develop intelligent systems for medical diagnostic support (CAD - Computer-Aided Diagnosis) and also significantly contribute to solving this very important problem.

The last dozen or so years have mainly seen attempts to determine quantitative parameters of pathologies occurring in the coronary vascularisation using various

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 193–202, 2011.

techniques, e.g. neural networks [6], wavelet-based fuzzy neural networks [7], Bayesian methods [8]. In this context one cannot find solutions in the form of intelligent systems that could imitate the thought processes taking place in the mind of the diagnosing physicians and thus generate a specific diagnosis, and not just a quantitative assessment of pathologies present. The attempt to solve a problem posed like this requires the use of advanced algorithms and information techniques which would help to penetrate the semantics of the image and as a result determine the appropriate therapeutic premises. The problem presented is complicated. Such a system cannot be developed using classical methods of the traditional automatic recognition and classification [13] – it is necessary to make use of a new technique of their automatic understanding leading to formulating semantic descriptions of analysed images [10]. Methods presented here refer to earlier publications by the authors and use the concept of the automatic understanding of medical images. The authors' success in developing tools for automatically understanding flat images [9], [10], [11] has persuaded them to propose similar methods for 3D representations of biological structures provided by modern imaging systems. Considerations presented herein will particularly apply to 3D visualizations of coronary vessels, which will form the foundation of the detailed part of this publication, but the formulation of the problem itself and the solution methodology outlined here can have much wider application.

2 Graph-Based Formalisms in Modeling, Semantic Description and Analysis for CT Visualizations of the Coronary Arteries

This section presents methods for making semantic models of 3D reconstructions of coronary arteries and interpretation capabilities based on these models.

2.1 Characteristics of the Set of Image Data

The work was conducted on images from diagnostic examinations made using 64-slice spiral computed tomography in the form of animations saved as AVI (MPEG4) files with the 512x512 pixel format. The analyzed images were acquired with a SOMATOM tomograph, which offers a number of opportunities to acquire data and make 3D reconstructions. Here it should be noted that our work omits the pre-processing stage, as coronary arteries are extracted using dedicated software integrated with the tomograph. In this software, predefined procedures have been implemented which allow the vascularisation to be quickly extracted from the visible structures of the cardiac muscle. Thus high quality images showing coronary arteries of the examined patient without ancillary elements were acquired. Consequently, this preprocessing stage was omitted, which allowed us to focus directly on the arteries examined. Since image data has been saved in the form of animations showing coronary vessels in various projections, for the further analysis we should select the appropriate projection which will show the examined coronary vessels in the most transparent form most convenient for describing and interpreting. In the clinical practice, this operation is done manually by the operator, who uses his/her own criteria to select the appropriate projection which shows the coronary vessels including their possible lesions. In our research we have attempted to automate the procedure of finding such

a projection by using selected geometric transformations during image processing. Using the fact that the spatial layout of an object can be determined by projecting it onto the axes of the Carthesian coordinate system, values of horizontal Feret diameters [14], which are a measure of the horizontal extent of the diagnosed coronary artery tree, are calculated for every subsequent animation frame during the image rotation. Determining these values is not particularly difficult and consists in calculating the difference between the maximum and the minimum coordinate of all points belonging to the projection of the analysed image on the X-axis. The projection for which the horizontal Feret diameter is the greatest is selected for further analyses, as this visualization shows both the right and the left coronary artery in the most convenient take. In a small number of analysed images, regardless of selecting the projection with the longest horizontal Feret diameter, vessels may obscure one another in space, which causes a problem at subsequent stages of the analysis. The best method to avoid this would be to use advanced techniques for determining mutually corresponding elements for every subsequent animation frame based on the geometric relations in 3D space [1], [2], [3].

2.2 Linguistic Formalisms in the Form of Indexed Edge-Unambiguous (IE) Graphs in 3D Modeling of Coronary Arteries

To help represent the examined structure of coronary vascularisation with a graph, it is necessary to define primary components of the analyzed image and their spatial relations, which will serve to extract and suitably represent the morphological characteristics significant for understanding the pathology shown in the image [15], [17]. It is therefore necessary to identify individual coronary arteries and their mutual spatial relations. To ease this process, the projection selected for analyzing was skeletonised. This made it possible to obtain the centre lines of examined arteries. These lines are equidistant from their external edges and one unit wide. This gives us the skeleton of the given artery which is much smaller than the artery itself, but fully reflects its topological structure. Skeletonising is aimed only at making possible to find bifurcation points in the vascularisation structures, and then to introduce an unambiguous linguistic description for individual coronary arteries and their branches. The morphometric parameters have to be determined based on a pattern showing the appropriate vessel, and not just only its skeleton. Of several skeletonising algorithms used to analyse medical images, the Pavlidis skeletonising algorithm [16] turned out to be one of the best. It facilitates generating regular, continuous skeletons with a central location and one unit width. It also leaves the fewest apparent side branches in the skeleton and the lines generated during the analysis are only negligibly shortened at their ends. The centre lines of analyzed arteries produced by skeletonising them is then searched for informative points, i.e. points where artery sections intersect or end. These points will constitute the vertices of a graph modeling the spatial structure of the coronary vessels of the heart. The next step is labelling them by giving each located informative point the appropriate label from the set of node labels. In the case of terminal points (leaves of a graph modeling the coronary vascularisation), the set of node labels comprises abbreviated names of arteries found in coronary vascularisation. They have been defined as follows: for the left coronary artery: LCA - left coronary artery, LAD - anterior interventricular branch (left anterior descending), CX - circumflex branch,

L - lateral branch, LM - left marginal branch and for the right coronary artery: RCA - right coronary artery, A - atrial branch, RM - right marginal branch, PI - posterior interventricular branch, RP - right posterolateral branch. This way, all initial and final points of coronary vessels as well as all points where main vessels branch or change into lower level vessels have been determined and labelled as appropriate. After this operation, the coronary vascularisation tree is divided into sections which constitute the edges of a graph modeling the examined coronary arteries. This makes it possible to formulate a description in the form of edge labels which determine the mutual spatial relations between the primary components, i.e. between subsequent arteries shown in the analysed image. These labels have been identified according to the following system. Mutual spatial relations that may occur between elements of the vascular structure represented by a graph are described by the set of edges. The elements of this set have been defined by introducing the appropriate spatial relations: vertical - defined by the set of labels {α, β,…, μ} and horizontal - defined by the set of labels {1, 2,…, 24} on a hypothetical sphere surrounding the heart muscle. These labels designate individual final intervals, each of which has the angular spread of 15°. Then, depending on the location, terminal edge labels are assigned to all branches identified by the beginnings and ends of the appropriate sections of coronary arteries. The use of the presented methodology to determine spatial relations for the analysed projection is shown below (fig. 1).

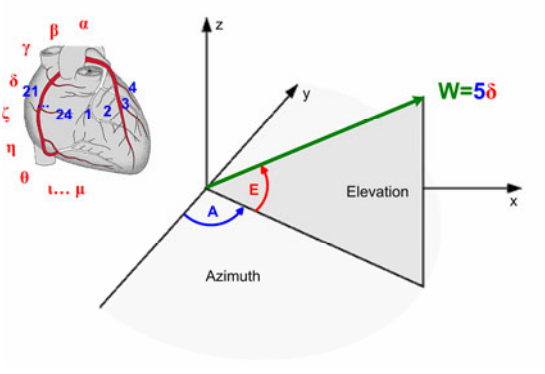
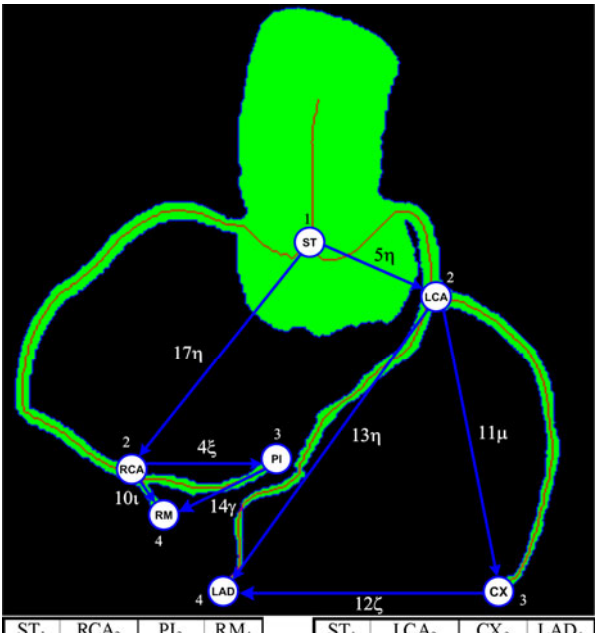


Fig. 1. Procedure of identifying spatial relations between individual coronary arteries

This representation of mutual spatial relations between the analysed arteries yields a convenient access to and a unanimous description of all elements of the vascular structure. At subsequent analysis stages, this description will be correctly formalised using ETPL(k) (Embedding Transformation-preserved Production-ordered k-Left nodes unambiguous) graph grammars defined in [18], supporting the search for stenoses in the lumen of arteries forming parts of the coronary vascularisation. ETPL(k) grammars generate IE graphs which can unambiguously represent 3D structures of heart muscle vascularisation visualized in images acquired during diagnostic examinations with the use of spiral computed tomography. Before we define the

representation of the analysed image in the form of IE graphs, we have to introduce the following order relationship in the set of Γ edge labels: $1 \leq 2 \leq 3 \leq \ldots \leq 24$ and $\alpha \leq \beta \leq \gamma \leq \ldots \leq \mu$. This way, we index all vertices according to the $\leq$ relationship in the set of edge labels which connect the main vertex marked 1 to the adjacent vertices and we index in the ascending order (i = 2, 3, …, n). After this operation every vertex of the graph is unambiguously assigned the appropriate index which will later be used when syntactically analysing the graph representations examined. IE graphs generated using the presented methodology, modeling the analysed coronary vascularisation, are presented in the figure below (fig. 2).



ST_1	RCA_2	PI_3	RM_4
1	2	1	–
17η	4ξ 10ι	14γ	–
2	3 4	4	–

ST_1	LCA_2	CX_3	LAD_4
1	2	1	–
5η	11μ 13η	12ζ	–
2	3 4	4	–

Fig. 2. The representation of the right and the left coronary artery using IE graphs. Their characteristic descriptions are presented in the tables below the figure.

The graph structure created in this way will form elements of a graph language defining the spatial topology of the heart muscle vascularisation including its possible morphological changes. Formulating a linguistic description for the purpose of determining the semantics of the lesions searched for and identifying (locating) pathological stenoses will support the computer analysis of the structure obtained in order to automatically detect the number of stenoses, their location, type (concentric or eccentric) and extent. For IE graphs defined as above, in order to locate the place where stenoses occur in the case of a balanced artery distribution, the graph grammar may take the following form:

a) for the right coronary artery:

$$G_R = (\Sigma, \Delta, \Gamma, P, Z) \tag{1}$$

Σ = {ST, RCA, RM, PI, C_Right}
Δ = {ST, RCA, RM, PI}
Γ = {17η, 4ξ, 10ι, 14γ}
Z is the start graph (fig. 3.)
P is the set of productions shown in fig. 3.

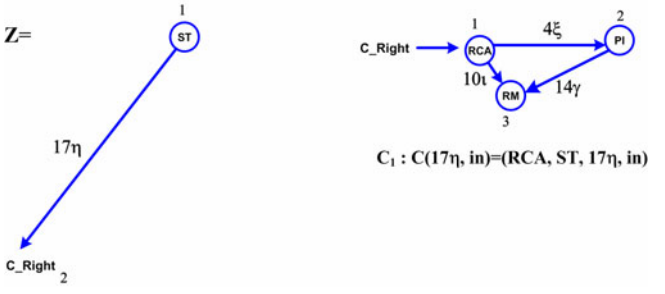


Fig. 3. Set of productions for grammar G_R

b) for the left coronary artery:

$$G_L = (\Sigma, \Delta, \Gamma, P, Z) \tag{2}$$

Σ = {ST, LCA, CX, LAD, C_Left}
Δ = {ST, LCA, CX, LAD}
Γ = {5η, 11μ, 13η, 12ξ}
Z is the start graph (fig. 4.)
P is the set of productions shown in fig. 4.

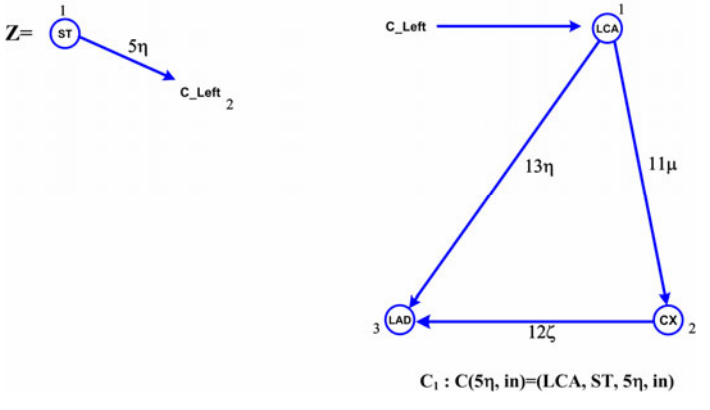


Fig. 4. Set of productions for grammar G_L

This way, we have defined a mechanism in the form of ETPL(k) graph grammars which create a certain linguistic representation of each analysed image in the form of an IE graphs. The set of all representations of images generated by this grammar is treated as a certain language. Consequently, we can built a syntax analyser based on the proposed graph grammar which will recognise elements of this language. The syntax analyser is the proper program which will recognise the changes looked for in the lumen of coronary arteries.

2.3 Constructing the Syntactic Analyser and Analysis of Particular Sections of Coronary Arteries

Defining the appropriate mechanism in the form of a G graph grammar yielded IE graph representations for the analysed images, and the set of all image representations generated by this grammar is treated as a certain language. The most important stage which corresponds to the recognition procedure for the pathologies found is the implementation of a syntactic analyser which would allow the analysis to be carried out using cognitive resonance [10], [12], which is key to understanding the image. One of the procedures of algorithms for parsing IE graphs for ETPL(k) graph grammars is the comparison of descriptions of characteristic subsequent vertices of the analysed IE graph and the derived IE graph which is to lead to generating the analysed graph. A one-pass generation-type parser carries out the syntactic analysis, at every step examining the characteristic description of the given vertex. If the vertex of the analysed and the derived graphs is terminal, then their characteristic descriptions are examined. However, if the vertex of the derived graph is non-terminal, then a production is searched for, after the application of which the characteristic descriptions are consistent. The numbers of productions used during the parsing form the basis for classifying the recognised structure. This methodology makes use of theoretical aspects of conducting the syntactic analysis for ETPL(k) grammars, described in [17], [18], [19]. Because of practical applications, the graph grammar should facilitate effective parsing. This has been achieved by using ETPL(k) grammars. For such grammars, there are deterministic automata supporting effective parsing with the computational complexity of $O(n^2)$ [17], [18], [19]. The ability to carry out semantic actions assigned to individual productions makes it possible to generate certain values or information which results from the syntactic analysis completed. In the case of analyses of 3D coronary vascularisation reconstructions, semantic actions of some productions will be aimed at determining numerical values defining the location and degree of the stenosis as well as its type (concentric or eccentric). These parameters will then be used as additional information useful in recognising doubtful or ambiguous cases or symptoms.

3 Results and Discussion

In order to determine the operating efficiency of the proposed methods, a set of test data was used. This set included 20 complete reconstructions of coronary

vascularisation obtained during diagnostic examinations using 64-slice spiral CT for various patients, and consist three types of topologies of coronary vascularisation (i.e. balanced artery distribution, right artery dominant, left artery dominant). The test data also included visualizations previously used to construct the grammar and the syntactic analyser. However, to avoid analysing identical images from the same sequences we selected frames that were several frames later than the projections used to construct the set of grammatical rules, and we used these later frames for the analysis. The above set of image data was used to determine the percentage efficiency of correct recognitions of the stenoses present, using the methodology proposed here. The recognition consists in identifying the locations of stenoses, their number, extent and type (concentric or eccentric). For the research data included in the experiment, 85% of recognitions were correct. This value is the percentage proportion of the number of images in which the occurring stenoses were correctly located, measured and properly interpreted to the number of all analysed images included in the experimental data set. The error percentage 15% of the experimental results, obtained for the proposed methods, concern problems in the qualifying stage of the image to the using an appropriate one of the defined graph grammars, which were created for different types of vascular topology. This was the case with the images on which the structure of coronary vascularisation contained no significant features belonging to the one of three distinguished sets. In order to assess whether the size of the stenosis was correctly measured using the semantic actions defined in the grammar, we used comparative values from the syngo Vessel View software forming part of the HeartView CI suite [20]. This program is used in everyday clinical practice where examinations are made with the SOMATOM Sensation Cardiac 64 tomograph. In order to confirm or reject the regularity of the stenosis type determination (concentric or eccentric) shown in the examined image, we decided to use a visual assessment, because the aforementioned programs did not have this functionality implemented. As the set of test data was small (20 elements) the results obtained are very promising and show that graph languages for describing shape features can be effectively used to describe 3D reconstructions of coronary vessels and also to formulate semantic meaning descriptions of lesions found in these reconstructions. Such formalisms, due to their significant descriptive power (characteristic especially for graph grammars) can create models of both examined vessels whose morphology shows no lesions and those with visible lesions bearing witness to early or more advanced stages of the ischemic heart disease. An additional advantage of the discussed image languages which use advanced formalisms of mathematical linguistics, translation theory and graph theory, is that they automatically identify significant informative points on the image which indicate the presence of lesions. In addition, by introducing the appropriate spatial relations into the coronary vessel reconstruction, it is possible to reproduce their biological role, namely the blood distribution within the whole coronary circulation system, which also facilitates locating and determining the progression stage of lesions. All of this makes up a process of automatically understanding the examined 3D structure, which allows us to provide the physician with far more and far more valuable

premises for his/her therapeutic decisions than we could if we were using the traditional image recognition paradigm. The other possible application areas of the proposed methods such another organs 3D image diagnosis are diverse. However, it must be possible to introduce graph representation of the visualized organ, which in turn will allow to define the appropriate graph grammar, enabling the representation of the image together with the potential morphological changes. Future works will include, among others problems related to automating the process of generating new grammars for cases not included in the present language. Another planned element of further research is to focus on using linguistic artificial intelligence methods to create additional, effective mechanisms which can be used for indexing and quickly finding specialised image data in medical databases.

Acknowledgments. This work has been supported by the Ministry of Science and Higher Education, Republic of Poland, under project number N N516 478940.

References

1. Lewandowski, P., Tomczyk, A., Szczepaniak, P.S.: Visualization of 3-D Objects in Medicine - Selected Technical Aspects for Physicians. Journal of Medical Informatics and Technologies 11, 59–67 (2007)
2. Higgins, W.E., Reinhardt, J.M.: Cardiac image processing. In: Bovik, A. (ed.) Handbook of Video and Image Processing, pp. 789–804. Academic Press, London (2000)
3. Sonka, M., Fitzpatrick, J.M.: Handbook of Medical Imaging. Medical Image Processing and Analysis, vol. 2. SPIE, Belligham (2004)
4. Wang, Y., Liatsis, P.: A Fully Automated Framework for Segmentation and Stenosis Quantification of Coronary Arteries in 3D CTA Imaging, dese. In: 2009 Second International Conference on Developments in eSystems Engineering, pp. 136–140 (2009)
5. Oncel, D., Oncel, G., Tastan, A., Tamci, B.: Detection of significant coronary artery stenosis with 64-section MDCT angiography. European Journal of Radiology 62(3), 394–405 (2007)
6. Cios, K.J., Goodenday, L.S., Merhi, M., Langenderfer, R.A.: Neural networks in detection of coronary artery disease. In: Proc. Computers in Cardiology, pp. 33–37 (1990)
7. Akay, Y.M., Akay, M., Welkowitz, W., Kostis, J.: Noninvasive detection of coronary artery disease using wavelet based fuzzy neural networks. IEEE EMB. Mag. 13(5), 761–764 (1994)
8. Cios, K.J., Goodenday, L.S., Wedding, A.: A Bayesian approach for dealing with uncertainties in detection of coronary artery stenosis using a knowledge-based system, DK. IEEE Eng. Med. Biol. Mag. 8(4), 53–58 (1989)
9. Ogiela, M.R., Tadeusiewicz, R.: Syntactic reasoning and pattern recognition for analysis of coronary artery images. Artificial Intelligence in Medicine 26, 145–159 (2002)
10. Tadeusiewicz, R., Ogiela, M.R.: Medical Image Understanding Technology. Springer, Heidelberg (2004)
11. Tadeusiewicz, R., Ogiela, M.R.: Structural Approach to Medical Image Understanding. Bulletin of the Polish Academy of Sciences – Technical Sciences 52(2), 131–139 (2004)
12. Ogiela, M.R., Tadeusiewicz, R.: Modern Computational Intelligence Methods for the Interpretation of Medical Images. Springer, Berlin (2008)

13. Meyer-Baese, A.: Pattern Recognition in Medical Imaging. Elsevier-Academic Press, Amsterdam (2003)
14. Tadeusiewicz, R., Korohoda, P.: Computer Analysis and Image Processing. Foundation of Progress in Telecommunication, Cracow (1997) (in Polish)
15. Tanaka, E.: Theoretical aspects of syntactic pattern recognition. Pattern Recognition 28, 1053–1061 (1995)
16. Pavlidis, T.: Algorithms for graphics and image processing. Computer Science Press, Rockville (1982)
17. Tadeusiewicz, R., Flasiński, M.: Pattern Recognition. PWN, Warsaw (1991) (in Polish)
18. Flasiński, M.: On the parsing of deterministic graph languages for syntactic pattern recognition. Pattern Recognition 26, 1–16 (1993)
19. Skomorowski, M.: A Syntactic-Statistical Approach to Recognition of Distorted Patterns. Jagiellonian University, Krakow (2000)
20. Get the Entire Picture, SOMATOM Sensation Cardiac 64 Brochure, Siemens medical (2004)
21. World Health Organization (WHO), The top ten causes of death - Fact sheet N310 (October 2008)

Discriminant Orthogonal Rank-One Tensor Projections for Face Recognition

Chang Liu[1,2,3], Kun He[3], Ji-liu Zhou[3], and Chao-Bang Gao[1,2,3]

[1] College of Information Science and Technology, Chengdu University, Chengdu, China
[2] Key Laboratory of Pattern Recognition and Intelligent Information Processing of Sichuan, Chengdu, China
[3] School of Computer Science, Sichuan University, Chengdu, China
chang.liu.scu@gmail.com, {hekun,zhoujl}@scu.edu.cn, kobren427@163.com

Abstract. Traditional face recognition algorithms are mostly based on vector space. These algorithms result in the curse of dimensionality and the small-size sample problem easily. In order to overcome these problems, a new discriminant orthogonal rank-one tensor projections algorithm is proposed. The algorithm with tensor representation projects tensor data into vector features in the orthogonal space using rank-one projections and improves the class separability with the discriminant constraint. Moreover, the algorithm employs the alternative iteration scheme instead of the heuristic algorithm and guarantees the orthogonality of rank-one projections. The experiments indicate that the algorithm proposed in the paper has better performance for face recognition.

Keywords: Discriminant constraint; Tensor representation; Orthogonal tensor, Rank-one projection; Face recognition.

1 Introduction

Many computer vision applications require processing a large amount of multi-dimension data, such as image, video data. Traditional dimensionality reduction algorithms generally transform each multi-dimension data into a vector by concatenating rows, which is called vectorization. Such kind of vectorization largely increases the computational cost of data analysis and seriously destroys the intrinsic tensor structure of high-order data. Moreover, these vector-based dimensionality reduction algorithms will over-fit the data for the small-size training set.

To address the problems caused by vectorization, [1]firstly has introduced tensor algebra into computer vision. Subsequently, there has been a growing interest in the development of supervised tensor dimensionality reduction algorithms [2,3,4,5]. [6] maximizes the trace ratio of inter-class scatter and intra-class scatter in the tensor metric, and proposes a Multi-linear Discriminant Analysis (MDA) algorithm. [7] proposes a General Tensor Discriminant Analysis (GTDA) algorithm based on Differential Scatter Discriminant Criterion (DSDC). MDA

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 203–211, 2011.

and GTDA project original tensor data into low-dimensional tensor features, i.e. Tensor-Tensor Projection (TTP). Different from TTP, [8] develops Tensor Rank-one Discriminant Analysis (TR1DA) based on Tensor-Vector Projection (TVP). TR1DA uses DSDC to obtain a number of rank-one projections from the repeatedly-calculated residues of the original tensor data. However, it is difficult to choose the optical weight of the intra-class scatter in TR1DA and the "greedy" approach used in TR1DA is a heuristic method which results in high computation cost. An Uncorrelated Multi-linear Discriminant Analysis (UMLDA) algorithm is proposed in [9] to extract uncorrelated features through TVP. [10] conducts a large number of face recognition experiments and indicates that the tensor-based scatter ratio criterion is superior to DSDC and TVP has better performance than TTP.

Recent researches [11,12]point out that enforcing an orthogonality constraint between the projection directions is more effective for preserving the intrinsic geometrical structure of original data. Consequently, the paper develops a Discriminant Orthogonal Rank-one Tensor Projections algorithm (DOR1TP). The algorithm uses tensor-based scatter ratio criterion to obtain orthogonal rank-one tensor projections directly form tensor data through solving TVP. The solution consists of sequential iterative processes based on the alternative iterative scheme. For each iterative, inter-class scatter and intra-class scatter is updated, which is effectively to enhance the class separability power of the low-dimensional features. The face recognition experiments show that DOR1TP outperform two vector-based algorithms (ULDA, OLDA) and four tensor-based algorithms (MDA, TR1DA, GTDA, UMLDA).

The remainder of the paper is organized as follows:Section 2 formulates the problem of DOR1TP and derives the solution process using the alternative iterative scheme. In Section3, the experiments on two face databases are reported, and the detailed recognition results are then compared against competing vector-based algorithms as well as tensor-based algorithms. Finally conclusions are drawn in Section 4.

2 Discriminant Orthogonal Rank-One Tensor Projections

Giving a set of training samples in high-dimensional space $X = \{X_i \in \mathbb{R}^{I_1 \times I_2 \times \cdots I_N}, i = 1, 2, ..., M\}$ with class labels $C = \{c_1, c_2, ...c_h\}$, where each sample $X_m \in \mathbb{R}^{I_1 \times I_2 \times I_3 \ldots \times I_N}$ is a N-order tensor, h denotes the class number. Based on TVP, original tensor data is projected into the low-dimensional vector space, where $Y_i \in \mathbb{R}^{P \times 1}$:

$$Y_i = X_i \times_{n=1}^{N} \{U_p^{(n)^T}, n = 1, ..., N\}_{p=1}^{P} \tag{1}$$

where $\left\| U_p^{(n)} \right\| = 1$, Let $Y_i(p)$ be the$p$th element of the vector Y_i, we have

$$Y_i(p) = X_i \times_1 U_p^{(1)^T} \times_2 U_p^{(2)^T} \cdots \times_N U_p^{(N)^T} \tag{2}$$

For pursuing a discriminative embedding for face recognition, we always hope to preserve the within-class neighborhood relationship while dissociating the sub-manifolds for different classes from each other. Furthermore, in order to preserve the intrinsic geometrical structure of original data, we also hope that the projection directions are orthogonal mutually. Therefore, our objective here is to learn a set of orthogonal rank-one tensor projections$\{U_p^{(n)} \in \mathbb{R}^{I_n}, n = 1, \cdots, N, p = 1, \cdots, P\}$such that the distances for those samples with the same class are minimized while the distances for those samples with different classes are maximized in the low-dimensional space. Consequently, we adopt the tensor-based scatter ratio criterion to project training sample X_i from tensor space $\mathbb{R}^{I_1} \otimes \cdots \otimes \mathbb{R}^{I_N}$ into low-dimensional vector space $Y_i \in \mathbb{R}^{P \times 1}$:

$$\arg\max \frac{S_b}{S_w} \qquad s.t. U_1 \perp U_2 \perp \cdots \perp U_P \tag{3}$$

where

$$S_w = \sum_{r=1}^{r=h} \sum_{Y_i \in c_r} (Y_i - m_r)(Y_i - m_r)^T \tag{4}$$

$$S_b = \sum_{r=1}^{h} N_r (m_r - m)(m_r - m)^T \tag{5}$$

$$Y_i = X_i \times_{n=1}^{N} \{U_p^{(n)^T}, n = 1, ..., N\}_{p=1}^{P} \tag{6}$$

in which N_r is the number of samples in the r class, m_r is the class mean of the low-dimensional features in the r class, m is the mean of all low-dimensional features. It follows that at least one pair vectors must satisfy $U_{p_1}^{(i)} \perp U_{p_2}^{(i)}$, if we require $U_{p_1} \perp U_{p_2}$,which means that rank-one tensors U_{p_1} and U_{p_2} are orthogonal.

According to the alternative iterative algorithm, we suppose $\{U_p^{(1)}, \cdots, U_p^{(n-1)}, U_p^{(n+1)} \cdots, U_p^{(N)}\}$ is fixed, the pth rank-one projection $\{U_p^{(1)}, \cdots, U_p^{(N)}\}$ is solved with the following optimization problem:

$$\arg\max_{U_p^{(n)}} \frac{U_p^{(n)^T} S_{bp}^{(n)} U_p^{(n)}}{U_p^{(n)^T} S_{wp}^{(n)} U_p^{(n)}} \tag{7}$$

where

$$S_{wp}^{(n)} = \sum_{r=1}^{h} \sum_{Y_i^{(n)} \in c_r} (Y_i^{(n)} - m_r^{(n)})(Y_i^{(n)} - m_r^{(n)})^T \tag{8}$$

$$S_{bp}^{(n)} = \sum_{r=1}^{h} N_r (m_r^{(n)} - m^{(n)})(m_r^{(n)} - m^{(n)})^T \tag{9}$$

$$Y_i = X_i \times U_p^{(1)^T} \times \cdots \times U_p^{(n-1)^T} \times U_p^{(n+1)^T} \times \cdots \times U_p^{(N)^T} \tag{10}$$

For ease of calculation, we assume that the kth dimension of all rank-one tensors are orthogonal, then the orthogonal constraint in Equation (3) is rewritten as:

$$U_1^{(k)} \perp U_2^{(k)} \perp \cdots \perp U_p^{(k)} \tag{11}$$

When $p = 1$, there is only one rank-one projection, it is straight to solve the unconstrained optimization problem of Equation (7). When $p > 1$,we need to consider two different situations. For$n \neq k$, we only need to solve the unconstrained optimization problem of Equation (7). But for $n = k$, we should to combine Equation (7) with Equation (11) into the following optimization problem:

$$\begin{aligned} &\arg\max_{U_p^{(k)}} \frac{U_p^{(k)^T} S_{bp}^{(k)} U_p^{(k)}}{U_p^{(k)^T} S_{wp}^{(k)} U_p^{(k)}} \\ &s.t. U_1^{(k)} \perp U_2^{(k)} \perp \cdots \perp U_p^{(k)} \end{aligned} \tag{12}$$

When $S_{wp}^{(k)}$ is non-singular, Equation (12) can be rewritten as follow:

$$\begin{aligned} &\arg\max_{U_p^{(k)}} U_p^{(k)^T} S_{bp}^{(k)} U_p^{(k)} \\ &s.t. U_p^{(k)^T} S_{wp}^{(k)} U_p^{(k)} = 1, U_p^{(k)^T} U_{p-1}^{(k)} = U_p^{(k)^T} U_{p-2}^{(k)} = ... = U_p^{(k)^T} U_1^{(k)} = 0 \end{aligned} \tag{13}$$

To solve the above constrained optimization problem, we formulate the following Lagrangian multipliers:

$$\begin{aligned} L = U_p^{(k)^T} S_{bp}^{(k)} U_p^{(k)} - \lambda (U_p^{(k)^T} S_{wp}^{(k)} U_p^{(k)} - 1) - \eta_{p-1} U_p^{(k)^T} U_{p-1}^{(k)} \\ -\eta_{p-2} U_p^{(k)^T} U_{p-2}^{(k)} - ... - \eta_1 U_p^{(k)^T} U_1^{(k)} \end{aligned} \tag{14}$$

Take the derivative of L with respect to $U_p^{(k)}$ and set it to zero, we have:

$$\frac{\partial L}{\partial U_p^{(k)}} = 2 S_{bp}^{(k)} U_p^{(k)} - 2\lambda S_{wp}^{(k)} U_p^{(k)} - \eta_{p-1} U_{p-1}^{(k)} - \eta_{p-2} U_{p-2}^{(k)} - ... - \eta_1 U_1^{(k)} = 0 \tag{15}$$

Left multiply both side of Equation (15) by $U_p^{(k)^T}$, and we can get:

$$2 U_p^{(k)^T} S_{bp}^{(k)} U_p^{(k)} - 2\lambda U_p^{(k)^T} S_{wp}^{(k)} U_p^{(k)} = 0 \tag{16}$$

Then we have:

$$\lambda = \frac{U_p^{(k)^T} S_{bp}^{(k)} U_p^{(k)}}{U_p^{(k)^T} S_{wp}^{(k)} U_p^{(k)}} \tag{17}$$

Therefore, λ is the objective value of Equation (7). Multiply both side of (15) by $U_q^{(k)^T} S_{wp}^{(k)^{-1}}$ for $q = 1, \cdots, p-1$:

$$\begin{aligned} &2 U_q^{(k)^T} S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} \\ &= \eta_{p-1} U_q^{(k)^T} S_{wp}^{(k)^{-1}} U_{p-1}^{(k)} + \eta_{p-2} U_q^{(k)^T} S_{wp}^{(k)^{-1}} U_{p-2}^{(k)} + ... + \eta_1 U_q^{(k)^T} S_{wp}^{(k)^{-1}} U_1^{(k)} \\ &= \sum_{j=1}^{p-1} \eta_j U_q^{(k)^T} S_{wp}^{(k)^{-1}} U_j^{(k)} \end{aligned} \tag{18}$$

With easy manipulation, we obtain a set of $p-1$ equations:

$$\begin{aligned} 2U_1^{(k)^T} S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} &= \sum_{j=1}^{p-1} \eta_j U_1^{(k)^T} S_{wp}^{(k)^{-1}} U_j^{(k)} \\ &\cdots \\ &\cdots \\ 2U_{p-1}^{(k)^T} S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} &= \sum_{j=1}^{p-1} \eta_j U_{p-1}^{(k)^T} S_{wp}^{(k)^{-1}} U_j^{(k)} \end{aligned} \tag{19}$$

We can write Equation (19) more concisely in matrix form as:

$$\begin{aligned} & 2 \begin{pmatrix} U_1^{(k)^T} \\ \vdots \\ U_{p-1}^{(k)^T} \end{pmatrix} S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} \\ = & \begin{pmatrix} U_1^{(k)^T} S_{wp}^{(k)^{-1}} U_1^{(k)}, U_1^{(k)^T} S_{wp}^{(k)^{-1}} U_2^{(k)}, \cdots, U_1^{(k)^T} S_{wp}^{(k)^{-1}} U_{p-1}^{(k)} \\ \vdots \\ U_{p-1}^{(k)^T} S_{wp}^{(k)^{-1}} U_1^{(k)}, U_{p-1}^{(k)^T} S_{wp}^{(k)^{-1}} U_2^{(k)}, \cdots, U_{p-1}^{(k)^T} S_{wp}^{(k)^{-1}} U_{p-1}^{(k)} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{p-1} \end{pmatrix} \end{aligned} \tag{20}$$

We can further simplify Equation (20) to be:

$$2 \begin{pmatrix} U_1^{(k)^T} \\ U_2^{(k)^T} \\ \vdots \\ U_{p-1}^{(k)^T} \end{pmatrix} S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} = \begin{pmatrix} U_1^{(k)^T} \\ U_2^{(k)^T} \\ \vdots \\ U_{p-1}^{(k)^T} \end{pmatrix} S_{wp}^{(k)^{-1}} \begin{pmatrix} U_1^{(k)} & U_2^{(k)} & \cdots U_{p-1}^{(k)} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{p-1} \end{pmatrix} \tag{21}$$

Let $U = \left(U_1^{(k)} U_2^{(k)} \cdots U_{p-1}^{(k)}\right)$,$\eta = (\eta_1 \ \eta_2 \ \cdots \ \eta_{p-1})^T$,Equation (21) can be rewritten as follow:

$$2U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} = U^T S_{wp}^{(k)^{-1}} U\eta \tag{22}$$

So, we have:

$$\eta = 2(U^T S_{wp}^{(k)^{-1}} U)^{-1} U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} \tag{23}$$

Multiply both side of Equation (15) by $S_{wp}^{(k)^{-1}}$, we have:

$$2S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} - 2\lambda U_p^{(k)} = S_{wp}^{(k)^{-1}} \sum_{j=1}^{p-1} \eta_j U_j^{(k)} \tag{24}$$

We rearrange it to be in the matrix form and get:

$$2S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} - 2\lambda U_p^{(k)} = S_{wp}^{(k)^{-1}} \left(U_1^{(k)} \cdots U_{p-1}^{(k)}\right) \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_{p-1} \end{pmatrix} = S_{wp}^{(k)^{-1}} U\eta \tag{25}$$

Embedding Equation (23) into Equation (25), we obtain:

$$S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} - \lambda U_p^{(k)} = S_{wp}^{(k)^{-1}} U (U^T S_{wp}^{(k)^{-1}} U)^{-1} U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)} U_p^{(k)} \quad (26)$$

Therefore, we have:

$$\left[S_{wp}^{(k)^{-1}} S_{bp}^{(k)} - S_{wp}^{(k)^{-1}} U (U^T S_{wp}^{(k)^{-1}} U)^{-1} U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)} \right] U_p^{(k)} = \lambda U_p^{(k)} \quad (27)$$

Let $M = S_{wp}^{(k)^{-1}} S_{bp}^{(k)} - S_{wp}^{(k)^{-1}} U (U^T S_{wp}^{(k)^{-1}} U)^{-1} U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)}$, according to Equation (17),*lambda* is exactly the quantity that we want to maximize, so the optimal $U^(k)_p$ is the eigenvector corresponding to the largest eigenvalue for the matrix M . It is deserve to pay attention that, when $S^(k)_{wp}$ and U are square matrices and non-singular, $S_{wp}^{(k)^{-1}} U (U^T S_{wp}^{(k)^{-1}} U)^{-1} U^T S_{wp}^{(k)^{-1}} S_{bp}^{(k)} = S_{wp}^{(k)^{-1}} S_{bp}^{(k)}$ and the left side of Equation (27) is zero. To avoid this case, we generally use the regular approach, and add a small regular parameter μI to $U^T S_{wp}^{(k)^{-1}} U$, where $I \in \mathbb{R}^{(p-1)\times(p-1)}$ is an identity matrix.

3 Experiments

In this section, we conduct face recognition on YaleB[13] and PIE[14] to evaluate the effectiveness of DOR1TP, and compare it with these vector-based algorithms ULDA[15]OLDA[16] and these tensor-based algorithms UMLDA[9],TR1DA[7], MDA[6], GTDA[7].

Fig. 1. Some samples from Yale B (middle row) and PIE (bottom row) face database

In all the experiments, all face images are normalized into the size of 32×32 . We randomly choose $L(= 5, 10)$ face images of each person for training and the rest for testing. The number of iterations for tensor-based algorithms is set to 10. The maximum number of features tested for ULDA and OLDA is $C-1$, where C is the number of classes in training. For TVP-based algorithms UMLDATR1DA and DOR1TP, up to 100 features are tested. For TTP-based algorithms MDA and GTDA, each dimension is varied from 1 to 32. Without loss of generality, we only test the dimension reduction case of $d_1 = d_2$ for simplification. For TR1DA, set $\zeta = 2$ for $L = 5$,$\zeta = 0.8$ for $L = 10$ .For UMLDA, we choose a fixed regularization parameter $\eta = 5 \times 10^{-4}$ empirically. The nearest neighbor (NN) classifier is used for final recognition and classification.

3.1 YaleB Face Database

The Yale B database contains 21888 face images of 38 people under 9 poses and 64 illumination conditions. We only choose all the 2414 frontal images as sub-database. The recognition rates are shown in Table 1 with different algorithms, and Fig. 2 shows the best recognition results with the corresponding low dimensions. It can be seen that MDA and UMLDA is superior to GTDA and TR1DA, and the recognition accuracy of TR1DA is unstable with different dimensions, which means that tensor-based scatter ratio achieves better results than DSDC. When the dimension is small, OLDA is competitive with ULDA, but the recognition accuracy of OLDA decreases rapidly with large dimensions. DOR1TP consistently outperforms all the other algorithms.

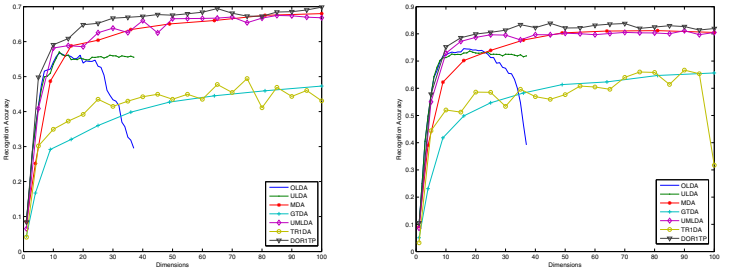


Fig. 2. (a)$L = 5$ (b)$L = 10$ The comparison of face recognition rates versus dimensions on YaleB

Table 1. Face recognition results on YaleB database

L	ULDA	OLDA	MDA	GTDA	TR1DA	UMLDA	DOR1TP
5	57.06%	56.70%	68.01%	53.01%	49.46%	67.48%	69.85%
10	73.70%	74.58%	83.33%	71.14%	66.67%	81.17%	83.88%

3.2 PIE Face Database

The PIE database includes 68 individuals with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. We choose 170 face images for each individual in our experiment. From Fig. 3 and Table 2, OLDA achieves better performance than ULDA and both of them become worse when the dimension is large. GTDA is worst, MDA and UMLDA are better than GTDA and TR1DA. DOR1TP still consistently outperforms other algorithms.

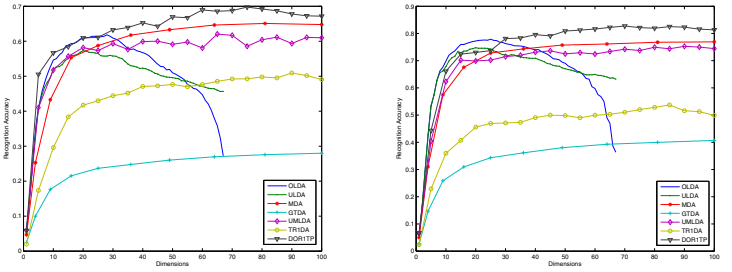


Fig. 3. (a)$L = 5$(b)$L = 10$ The comparison of face recognition rates versus dimensions on PIE

Table 2. Face recognition results on PIE database

L	ULDA	OLDA	MDA	GTDA	TR1DA	UMLDA	DOR1TP
5	57.31%	61.83%	65.69%	29.46%	50.93%	62.06%	69.76%
10	74.77%	77.90%	76.99%	43.61%	53.74%	75.26%	82.74%

4 Conclusions

The paper proposes a discriminant orthogonal rank-one tensor projections algorithm with the tensor-based scatter ratio criterion. The algorithm projects original tensor data into vector features using a set of orthogonal rank-one tensors, which combines the orthogonal constraint and the tensor-base scatter ratio criteria to improve the recognition accuracy. The face recognition experiments indicate that the algorithm proposed in the paper achieves better performance than other algorithms.

References

1. Vasilescu, M., Terzopoulos, D.: Multilinear analysis of image ensembles: TensorFaces. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 447–460. Springer, Heidelberg (2002)
2. Tao, D., Li, X., Wu, X., Hu, W., Maybank, S.: Supervised tensor learning. Knowledge and Information Systems 13, 1–42 (2007)
3. Lu, H., Plataniotis, K., Venetsanopoulos, A.: MPCA: Multilinear Principal Component Analysis of Tensor Objects. IEEE Transactions on Neural Networks 19, 18–39 (2008)
4. Zafeiriou, S.: Discriminant Nonnegative Tensor Factorization Algorithms. IEEE Transactions on Neural Networks 20, 217–235 (2009)
5. Lu, H., Plataniotis, K., Venetsanopoulos, A.: Gait recognition through MPCA plus LDA. In: Proceedings of the Biometric Consortium Conference, Baltimore, MD, USA, pp. 1–6 (2006)

6. Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X., Zhang, H.: Multilinear Discriminant Analysis for Face Recognition. IEEE Transactions On Image Processing 16, 212–220 (2007)
7. Tao, D., Li, X., Wu, X., Maybank, S.: General Tensor Discriminant Analysis and Gabor Features for Gait Recognition. IEEE Transactions On Pattern Analysis And Machine Intelligence 29, 1700–1715 (2007)
8. Tao, D., Li, X., Wu, X., Maybank, S.: Tensor Rank One Discriminant Analysis–A convergent method for discriminative multilinear subspace selection. Neurocomputing 71, 1866–1882 (2008)
9. Lu, H., Plataniotis, K., Venetsanopoulos, A.: Uncorrelated multilinear discriminant analysis with regularization and aggregation for tensor object recognition. IEEE Transactions On Neural Networks 20, 103–123 (2009)
10. Lu, H., Plataniotis, K., Venetsanopoulos, A.: A taxonomy of emerging multilinear discriminant analysis solutions for biometric signal recognition, Biometrics: Theory, Methods, and Applications, pp. 21–45. Wiley-IEEE (2009)
11. Kokiopoulou, E., Saad, Y.: Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 2143–2156 (2007)
12. Zhang, T., Huang, K., Li, X., Yang, J., Tao, D.: Discriminative orthogonal neighborhood-preserving projections for classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40, 253–263 (2010)
13. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 684–698 (2005)
14. Georghiades, A., Belhumeur, P., Kriegman, D.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Transactions on Pattern analysis and Machine Intelligence, 643–660 (2001)
15. Ye, J.: Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. Journal of Machine Learning Research 6, 483–502 (2005)
16. Nanni, L., Lumini, A.: Orthogonal linear discriminant analysis and feature selection for micro-array data classification. Expert Systems with Applications 37, 7132–7137 (2010)

Robust Visual Tracking Using Randomized Forest and Online Appearance Model

Nam Vo, Quang Tran, Thang Dinh, and Tien Dinh

Faculty of Information Technology, University of Science, VNU-HCMC
227 Nguyen Van Cu, Ho Chi Minh City, Vietnam
{nam.poke,tmquang}@gmail.com
{dbthang,dbtien}@fit.hcmus.edu.vn

Abstract. We propose a robust tracker based on tracking, learning and detection to follow an object in a long term. Our tracker consists of three different parts: a short term tracker, a detector, and an online object model. For the short-term tracker, we employ the Lucas Kanade tracker to keep following the object frame by frame. Meanwhile, the sequential randomized forest using a 5bit Haar-like Binary Pattern feature plays as a detector to detect all possible object candidates in the current frame. The online template-based object model consisting of positive and negative image patches decides which the best target is. Our method is consistent against challenges such as viewpoint changes, various lighting conditions, and cluttered background. Moreover, our method is efficiently able to reacquire the object efficiently even after it's out of view or in total occlusion. We also propose an efficient way to extend our tracker for multiple faces tracking application. Extensive experiments are provided to show the robust of our tracker. Comparisons with other state-of-the-art trackers are also demonstrated.

Keywords: long-term tracking, object detection, real-time.

1 Introduction

1.1 Motivation

Visual tracking is an active research task in computer vision with many real world applications such as video-surveillance, human-computer interaction, augmented-reality, content-based video retrieval, and robotics. The challenges in building a robust tracker are viewpoint changes, different lighting conditions, occlusions, and presence of cluttered background. Moreover, the tracking target may leave the field of view and reappear which requires a reliable appearance model in order to acquire it again.

1.2 Related Works

Tracking can be formulated as an estimation of the state in a time series. Given this formulation, probabilistic terms can be used. Assuming linear Gaussian distribution of the model, Kalman filter [4] was employed to seek for the optimal solution. Following

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 212–221, 2011.

the same formulation, Particle-filter [11], which estimates the state space by computing the posterior probability density function using Monte Carlo integration, is one of the most popular approaches. There are various variations and improvements developed from this original approach such as the hierarchical particle filter [13] and the cascade particle filter [14].

Recently, tracking-by-detection approach becomes popular in the research community. It considers tracking as a classification problem in which the classifier focus on separating the object from the background [5][6]. The online feature selection [7] is perhaps the pioneer in this trend. Meanwhile, Avidan *et. al.* proposed an integration of offline trained Support Vector Machine (SVM) classifier and a tracker using optical flow. Considering tracking as a semi-supervised problem, where the label data is the initial tagged object, and the unlabeled data is the incoming frame, Babenko *et. al.*[6] proposed to learn multiple instances of the object in an online manner so that the tracker can avoid the drifting problem. Generally, it is studied that building a robust appearance model is the vital part in visual tracking [12].

Adam *et. al.*[8] proposed to split the object into fragments to deal with partial occlusion. However, the method used a static appearance template which prevents it from dealing with complex background and appearance changes. Semi-boost [10], on the other hand, employed an online learned boosting classifier to find the decision boundary between the object and background. However, to avoid drifting, semi-boost requires the new object to have some similarity with the initial patch which is too conservative to adapt to appearance changes. Modeling the object appearance as a linear low-dimensional subspace, Ross *et. al.* [9] chose distance-to-subspace as the metric to decide the best target. Although this method is very adaptive to appearance changes, it is vulnerable to cluttered background with the limitation of discriminative powers. Trying to fuse the discriminative and generative models, Yu et. al. [12] proposed to use co-training framework; however, to decide the trade-off parameter between the two models is not a trivial work. More recently, Kalal *et. al.* [3] proposed an efficient method taking advantage of a sequential process of a tracker, a discriminative classifier, and a generative template-based model. This method is robust against complex background while adapting very well to appearance changes.

1.3 Our Approach

In this paper, inspired from the Tracking-Learning-Detection (TLD) Tracker [2][3], we propose a robust object method in the same fashion: combining a short-term tracker, an object detector, and an online model. Our contributions in this work are three-folds:

1) Extending the sequential randomized forest proposed in TLD for multiple object tracking. We also propose to use online color feature to narrow down the search space in order to speed up the whole system.
2) Introducing the new feature set, 5bitBP features, for efficient computation.
3) Proposing binary-search-tree (BST) structure to speed up the calculation process of the online model.

The rest of our paper is organized as follows. Our method is presented in section 2. Section 3 is the reported experiments followed by the conclusion in section 4.

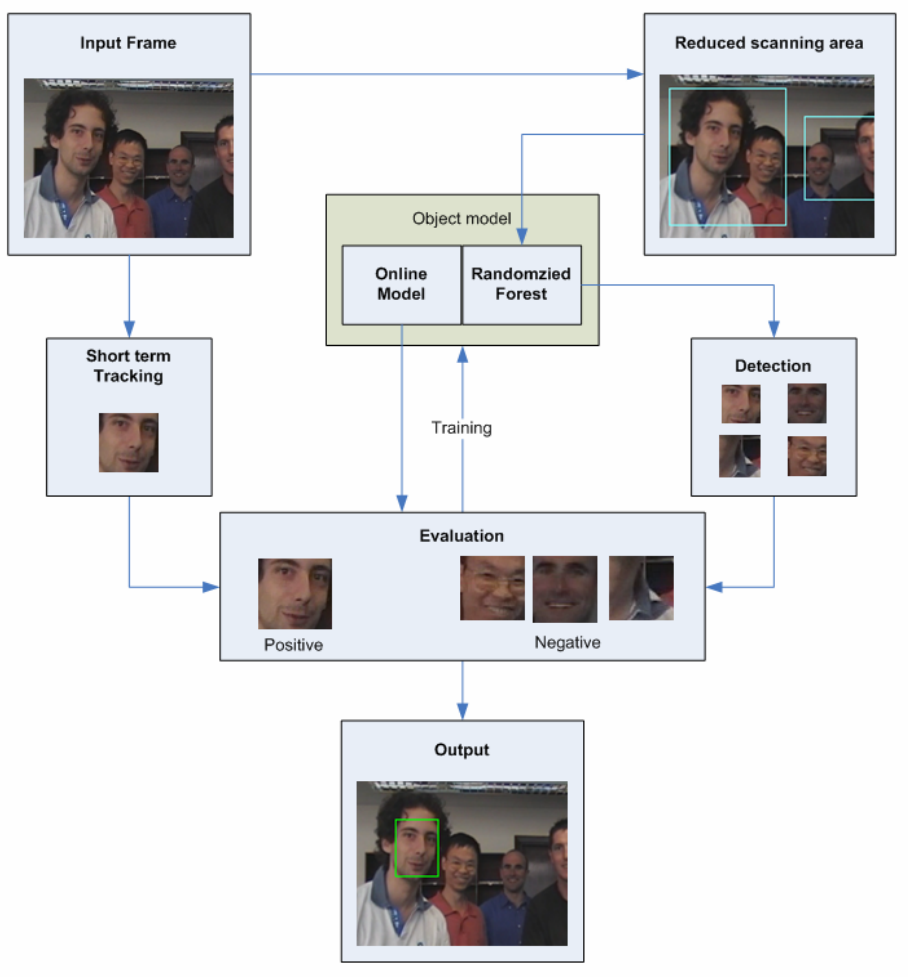


Fig. 1. An overview of our method

2 TLD – Tracking, Learning and Detection

The system consists of a short term tracker, a detector based on randomized forest, and an online model based on template matching. The combination of randomized forest and online model make up object appearance, the vital part of a tracker. The method works as follows (illustrated in Fig.1)

- The short term tracker tracks the object from the position in the previous frame. The Lucas Kanade Tracker is adopted for short term tracking.
- The randomized forest detector exhaustively scans the whole frame using the same strategy as in [15] and outputs all promising candidates. An efficient

online color feature selection algorithm can be applied to help reducing the scanning area.

- Among the candidates provided by the short term tracker and the detector, the best one decided by the online object model is the target. Others are used as negative samples for training both the randomized forest and the online model. Positive samples drawn randomly around the object position are also used for training both models.

2.1 Randomized Forest Detector with New 5-bit Haar-Like Binary Pattern Feature

The forest consists of a number of ferns. Each fern corresponds to a set of randomly generated Haar-like Binary Pattern feature. The feature value is discrete and each different combination of feature values corresponds to a leaf in the fern. The fern records the number of positive and negative samples (p and n respectively) added to its leaves during training. When checking a sample, we calculate its feature values to find the corresponding leaf in the fern. After that, the posterior is computed as $p / (p + n)$. The final posterior of the sample is the average output from all ferns. Samples with the final posterior of less than 50% are considered as background.

Similar to Haar-like feature [15], the Haar-like binary pattern feature, is used in the field of detection. The computation of theses kinds of features is often by comparing intensity average of pixels within a region to another region in the image (calling comparing regions for short).

Kalal *et. al.*[2] introduces the 2bit Binary Pattern (2bitBP) feature by comparing the average intensity between the half left region to the right left one, the half top region to the half bottom one in a predefined rectangle to extract a 2-bit value. Generally, a Haar-like binary pattern feature comparing all t regions with each other can produce up to $t!$ different codes, corresponding to $\lceil log2(t!) \rceil$ bit values. In other words, the number of comparisons needed to compute that feature is the same as sorting. Therefore, the compensation between the number of regions to be used and the comparison operations decides the efficiency in computation.

In this paper, we propose to use the 5 bit Haar-like Binary Pattern (5bitBP) feature set for efficient calculation. The four following regions are involved:

- A predefined rectangular region.
- The half left of the predefined rectangular region.
- The half top of the predefined rectangular region.
- The whole image patch. (note that this region is the same for all features; hence, it improves the time for computation)

This calculation produces 24 different codes which can be stored in a 5-bit integer number. Roughly, it takes 3 average calculations of region intensity and 5 comparison operations. Illustration of the 5bitBP feature is shown in Fig. 2.

In our implementation, the forest consists of 6 ferns; each fern makes use of 2 randomly generated 5bitBP features (e.g. 576 leaves/fern).

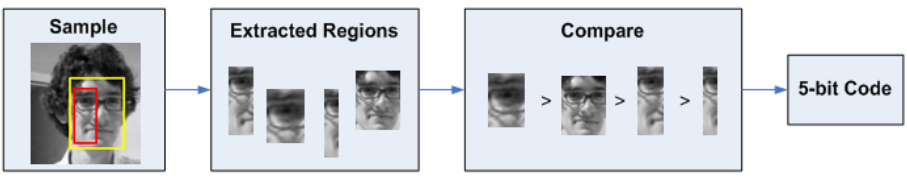


Fig. 2. The 5-bit haar-like Binary Pattern feature

2.2 Online Template-Based Model in a Binary Search Tree Structure

The online model stores image patches of positive and negative samples. To check a sample, we calculate its distances to positive and negative image patches in the online object model, denoted as d_p and d_n, respectively. The final distance is then computed as $d = d_p / (d_p + d_n)$.

The online model evaluates all object candidates collected by the randomized forest classifier and the short term tracker and outputs the one with shortest distance as target.

Distance measurement: distance between two image patches x_i, x_j is computed based on Normalized Cross-Correlation (NCC) operation:

$$distance(x_i, x_j) = 1 - NCC(x_i, x_j) \quad (1)$$

Note that for better performance the image patches should be first normalized by equalizing the histogram. To avoid the biased decision, the distance to positive/negative image patches is averaged from a number of closest ones. ($n_p = 3$, $n_n = 1$ distances for positive and negative, respectively).

Binary search tree structure: As the online model keeps accepting new training samples, the distance computational cost grows linearly with the number of image patches in the online model (i.e. $O(n)$). Because this operation is heavily used in the system, it significantly slows down the tracker after a matter of time. Observing that only several nearest image patches are involved in the distance computation process, we decided to organize image patches in the online model in BST fashion for faster access instead of using the simple brute force searching.

The k-means algorithm is used to divide set of sample image patches with the distance definition given above and the mean being the "center" sample (which has sum of distances from itself to all other samples in the set to be smallest).

The BST structure is constructed as following: dividing the set of sample image patches into 2 smaller set (using k-means with k = 2, also record the "center" of each set), if the sets are not small enough (e.g. less than $\theta=50$ in our implementation), continue dividing them.

After this process, a BST is built in which each node corresponds to a center, and each leaf corresponds to a small size set of image patches. A testing sample travels from the top of the tree until it reaches a leaf. At a node, the distances from that sample to the two centers are used to decide which path to take. The final distance then can be calculated using only image patches in the set of samples in the leaf it belongs to. This complexity of this process is $O(logn)$. Theoretically, the result might be suboptimal. However, in our experiments, comparable tracking results are achieved while the performance is significantly faster.

2.3 Training the Model

In the initialization step, the first object image patch, which is manually chosen, is added to the online model. A number of positive/negative samples are drawn randomly (near/far from the object) and added to the forest (which is added to all the ferns in the forest).

For each frame, the model is updated if the following conditions are satisfied:

- The object trajectory is validated: the object trajectory is validated when the distance of the last sample in the trajectory is smaller than a defined threshold d_{track}, or the distances between consecutive samples in trajectory are smaller than a defined threshold (d_{step} = 0.2). This accepted trajectory enables modeling the object appearance changes effectively.
- The object is still tracked: If the distance of the tracked object is larger than d_{lost} =0.5, the object is lost. It is tracked again whenever the distance of the output is smaller than d_{track} =0.35. This procedure ensures that the model is only updated when the object is tracked correctly.

During the update process, all samples are added to the forest. Then the samples in the validated trajectory are added to the online model if its distances to all positive image patches larger than a defined threshold ($d_{positive\_add}$ = 0.1). The reason is to prevent adding too many similar image patches to the online model allowing us to keep only the different representative poses of the object. The negative samples are added to the online model in the same manner ($d_{negative\_add}$ = 0.1).

Note that a pose of an object is usually close to some other poses of the same object. Thus, we define the distance of the sample to the positive image patches in the online model as the average distance to n_p closest ones (with n_p > 1). This measurement helps to resist drifting or false positive samples which are possibly similar to only a single pose of the object.

2.4 Scan Area Limitation Based on Color Information

Exhaustive scanning window technique is very expensive (as in original TLD [3]). However, it is possible to reduce the scanning area from the whole image to some regions with high confidence using color information, which is overlooked by the original TLD tracker. The illustration is shown in Fig. 3. Based on color of the object and the background in the previous frames, a likelihood image of current frame is constructed indicating some regions of interest. The process of scanning for object candidates is now narrowed down to these areas.

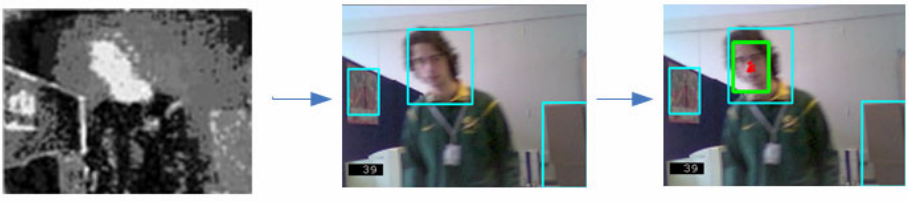


Fig. 3. The likelihood image reducing scanning area

Here, we applied the method described in [7] in order to best distinguish object and its surrounding area. In this technique, a set of features is chosen by linear combination of R, G, B color with relative weights. Each feature is scored using the variance ratio to measure how discriminative it is. We suggest using the "running average" method to update the scores (with learning rate $\alpha = 0.1$) and the feature with highest one is chosen to build the likelihood image.

$$var_{update} = \alpha\, var_{new} + (1-\alpha)\, var_{old} \tag{2}$$

2.5 Extension to Multi Face Tracking

In this section, we propose a way to extend our system for multi face tracking. A face detector [15] is used, periodically detects new face appearing in the scene and adds to the system.

In fact, multi object tracking can be done by using several independent single object trackers at the same time. However the approach is inefficient because all single object trackers have to scan the same frame. Our proposed multi face tracking system consists of single object trackers whose randomized forest share the same feature set; hence the process of scanning the frame calculating feature, which is the most expensive one, is only performed once and made use by all single object trackers.

Tracking multiple objects has the advantage of avoiding confusion while determining the right objects with similar appearance. Also, objects with discriminative intensity pattern like faces are very suitable to be tracked by TLD. Moreover, particular color of human skin can be effectively made use by technique mentioned in 2.4. Hence our system is expected to run faster and more accurate, comparing with running several single object trackers at the same time.

3 Experimental Studies

We have tested our tracker on many challenging video sequences. Table 1 show results from 6 sequences: David indoor, Plush toy [9], Pedestrian [1], Tiger 1, Tiger 2, Coke can [6]. In comparison, we used several state-of-art approaches: the original TLD [3], FragTracker [8], MILTrack [6]. All algorithms are implemented in C/C++. We also fix all our parameters (whose values have been mentioned in previous sections), except the object minimum size used when scanning the image due to difference in size of video and object.

In most sequences, our tracker outperforms FragTracker and MILTrack. Some snapshots are shown in Fig. 4. Notably are sequences Tiger 1, Tiger 2 and David indoor. Sequence Tiger 1, Tiger 2 show toy tiger moving fast causing blur and frequently under partly occlusion; our system produces comparable results as MILTrack. In David indoor, a person moves from a completely dark room to a bright lobby. It is very challenging due to extreme illumination change. Frag trackers and MILTrack have to be initialized at 300^{th} frame while ours is able to track the face of the person from the very first frame where it is too dark even for human eye to recognize whether there is an object moving.

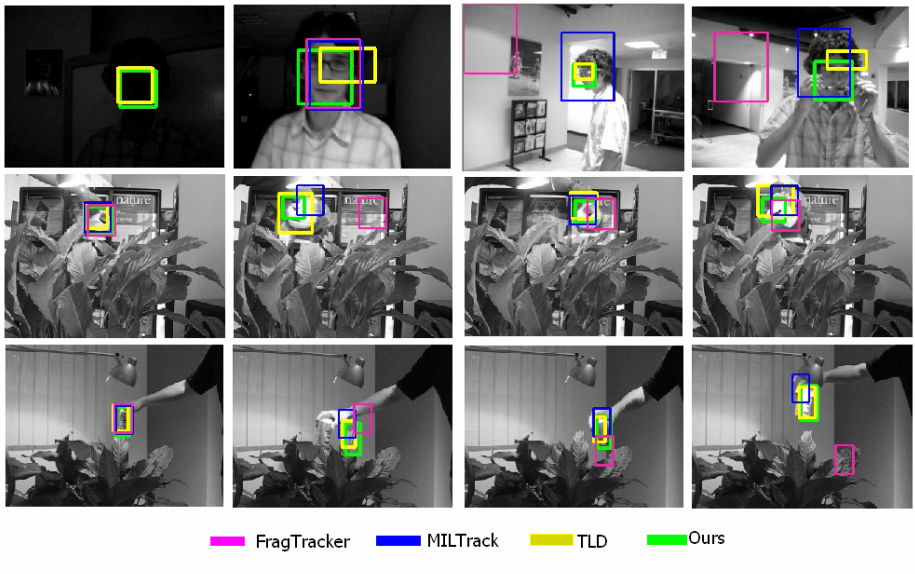


Fig. 4. Snapshots from *David indoor, Tiger 1*, *Coke can* sequences

Table 1 shows the comparison results using the average center location error metric [6] (average distance from object's center given by the tracker and one given by the groundtruth). It is important to emphasize that our system is at least twice faster than the original TLD while achieving comparative results. Processing a 320x240 sequence on the same 1.86GHz processor, our system runs at 15fps, while TLD runs at 6fps.

Table 1. Average center location error

Sequence	Frag	MILTrack	TLD	Ours
David indoor	23	46	14	5
Pedestrian	56	n/a	4	13
Plush toy	11	11	5	5
Tiger 1	40	15	13	17
Tiger 2	38	17	16	12
Coke can	63	21	9	9

To validate our multi face tracking system, we have tested on three multiple faces sequences from [17]: motinas_multi_face_frontal, motinas_multi_face_turning, motinas_multi_face_fast. The scenarios consist of 3, 4 targets moving around, repeatedly occluding each other while appearing and disappearing from the field of view (Fig. 5). Some targets are successfully tracked until the end with same identity; some are unable to recover after reappearing and are recognized as a new target and have different identities. It mainly happens in the beginning of the sequence when the poses of the object have not been fully learned by the tracker.

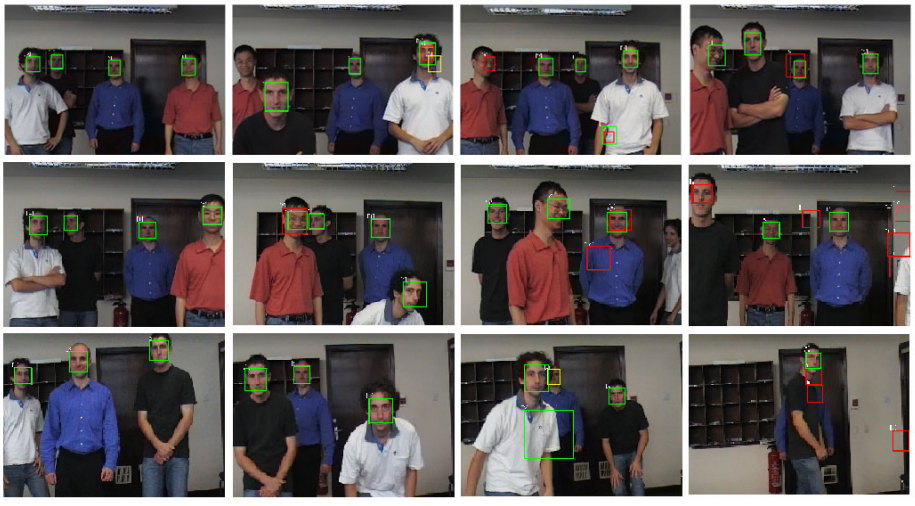


Fig. 5. Frames from output of motinas_multi_face_frontal, motinas_multi_face_turning and motinas_multi_face_fast sequence

4 Conclusion

We have presented a novel framework for object tracking. It is based on TLD concept which combines a short term tracker, a detector and an online object model. We have introduced the new feature set, 5bitBP feature, to improve the performance of the detector while proposing the BST structure for the online object model which outperforms the original brute forte technique. The efficient extension of the tracker to deal with multiple object scenarios has also been discussed. Extensive experiments and comparisons to other state-of-the art methods have been demonstrated the robustness of our tracker.

In the future, we would like to improve our multi object tracker system. We also plan to explore the context information to help the tracker avoid occlusion.

Acknowledgements

This work is a part of the KC.01/06-10 project supported by the Ministry of Science and Technology, 2009-2010.

References

1. Caviar Test Case Scenarios, `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/`
2. Kalal, Z., Matas, J., Mikolajczyk, K.: Online Learning of Robust Object Detectors During Unstable Tracking. In: OLCV (2009)

3. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In: CVPR (2010)
4. Kalman, R.E.: A New Approach to Linear Fltering and Prediction Problems. Transactons of the ACME-Journal of Basic Engineering, 35–45 (1960)
5. Avidan, S.: Support vector tracking. IEEE Trans. on PAMI, 184–191 (2001)
6. Babenko, B., Yang, M.-H., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. In: CVPR, pp. 983–990 (2009)
7. Collins, R., Liu, Y., Leordeanu, M.: Online Selection of Discriminative Tracking Features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1631–1643 (2005)
8. Adam, A., Rivlin, E., Shimshoni, I.: Robust Fragments-based Tracking Using The Integral Histogram. In: CVPR, vol. 1, pp. 798–805 (2006)
9. Ross, D., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental Learning for Robust Visual Tracking. In: IJCV, vol. 77, pp. 125–141 (2008)
10. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
11. Isard, M., Blake, A.: Condensation-Conditional Density Propagation for Visual Tracking. In: IJCV 1998, pp. 5–28 (1998)
12. Yu, Q., Dinh, T., Medioni, G.: Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 678–691. Springer, Heidelberg (2008)
13. Yang, C., Duraiswami, R., Davis, L.S.: Fast Multiple Object Tracking via A Hierarchical Particle Filter. In: ICCV, pp. 212–219 (2005)
14. Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in Low Frame Rate Video: A Cascade Particle Filter with Discriminative Observers of Different Lifespans. In: IEEE CVPR (2007)
15. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: CVPR (2001)
16. Maggio, E., Piccardo, E., Regazzoni, C., Cavallaro, A.: Particle PHD filter for Multi-target Visual Tracking. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, USA, April 15-20 (2007)

Graphical Pattern Identification Inspired by Perception

Urszula Markowska-Kaczmar and Adam Rybski

Wrocław University of Technology
urszula.markowska-kaczmar@pwr.wroc.pl

Abstract. The paper presents a method of static graphical pattern identification inspired by human perception. Cooperation of visual cortex regions is thought to play a major role in the perception, that is why in our approach this cooperation is modelled by joining algorithms responsible for shape and color processing. In order to obtain more stable set of characteristic points, the SIFT algorithm has been modified. To acquire information about shapes Harris operator and Hough transform are considered. The proposed method achieves about 28% less number of incorrect identification comparing to the results obtained by the classical SIFT algorithm.

Keywords: visual pattern identification, modified SIFT algorithm, shape analysis, perception.

1 Introduction

For a long time a great deal of research has been devoted to simulation of human behaviour. In visual pattern recognition many methods are also inspired by human cognitive processes. In this paper pattern recognition is assumed as an identification of objects in images on the basis of their form, color, texture and other attributes. Generally, the methods of pattern identification in images can be divided into three groups: based on templates, based on features matching, based on application of different kinds of neural networks.

In the templates based methods correlation algorithms can be mentioned. The idea behind them lies in comparison of images pixel by pixel. They use information contained in the image without preprocessing. An example is Haar classifier [9].

The second group methods use features that are stable independently of an observation angle. Such features are offered by the following well known methods: SIFT algorithm [6], [7], which examines characteristic gradient of colors, Harris operator which role is to search for corners and Hough transform that is used to isolate features of a particular shape within an image.

In the third group convolution neural network can be mentioned. It is applied to identification of objects in images [5]. As a second example in this group STA neural network [1] can be quoted. It is used to discover attention regions.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 222–231, 2011.

It is worth to mentioning about works that model some aspects of cognition in their approaches to identify objects in images. Very interesting one is presented by Wang [10]. The system called FVIP (Framework of Visual Information Processing) contains perception module and various kinds of memories where sensorial information is stored. Another example is the hybrid method described by Furesjo and others in [3], which joins several basic information about objects.

Neocognitron is also cognition based method which is a specific neural network. Its performance principle assumes that subsequent layers recognize more and more complex information on the basis of simpler information from the previous layer. The last layer gives an answer about identification of the pattern.

The agent techniques are also used to model cognitive approach to identify visual objects. The Learning Intelligent Distribution Agent (LIDA) [2] is one of the examples. Nowadays this system is intensively developed and examined.

In our study we focus on the more elementary process of cognition, namely perception. The aim of the research presented in the paper was development of a static visual object identification method based on this process. It seems natural that the system dedicated to identify objects is inspired by cognitive and brain performance. In order to evaluate the method efficiency it was implemented and tested. Because perception is an effect of many visual cortex regions cooperation, the method joins several algorithms which are responsible for delivering information about an object. The method is able to find a given pattern on the basis of its single visual representation. In other words, the method learns to recognize object on the basis of one exemplary pattern. The components of the method were chosen in such a way that the identification of patterns was resistant to various geometric transformations, partially covered pattern and changes in the illumination.

2 Description of the Proposed Method

However, nowadays we are not able to fully model human visual system but current physiology provides a description how the visual signal is processed and analysed. This knowledge was the basis for development of our system which we have called LLCPR (Low Level Cognitive Pattern Recognition).

During an image analysis human being does not process it as a whole, but considers and processes its various aspects. Generally, this information refers to: color, form and motion. LLCPR simulates two visual cortex regions - V4, which analyses colors and V3 which processes shapes because we have limited the class of images to static ones. Fig. 1 presents corresponding regions and elements of LLCPR system. Our assumption was to develop a method which is able to find the pattern in images on the basis of single graphical representation of this pattern and it should be robust on various kinds geometrical transformations, partial pattern covering and changes in image lightening. Taking into account these assumptions and variety of existing patterns we have limited the class of patterns being considered in our research to the set of objects which have

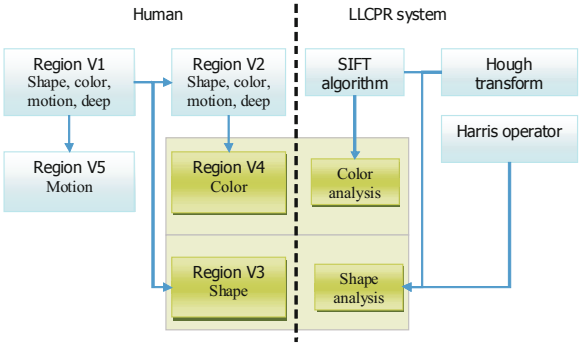


Fig. 1. Corresponding parts in human visual signals processing and LLCPR system

Fig. 2. Exemplary patterns considered in the experimental part

relatively easy geometrical structure. The examples of such objects are: game cards, company logos or traffic signs, some of them are shown in Fig. 2.

2.1 Color Information Acquisition and Comparison

Color information processing is based on SIFT algorithm. It extracts characteristic points of objects (patterns). Each point is represented by its location coordinates and represented by a feature vector built on the basis of color gradients between the point and its surroundings. Information about colors is called color descriptors. To recognize an object, a candidate matching between features is searched on the basis of Euclidian distance. In order to obtain more stable set of characteristic features, the SIFT algorithm has been modified. After assigning a set of characteristic points like in the classical SIFT algorithm, the final set of characteristic points is checked whether it satisfies particular conditions, i.e. it has to fulfill 3 criteria. They are briefly characterized below.

Distance criterion. For each characteristic point in the image the closest characteristic point in the pattern is searched. To find it the distance matrix $m \times n$ is built, where m is the number of points in the image, n is the number of points in the pattern.

$$M = \begin{bmatrix} d_{11}\ d_{12}\ \dots \\ d_{21}\ d_{22}\ \dots \\ \vdots\ \ \vdots\ \ \ddots \end{bmatrix} \tag{1}$$

In the matrix (1) the value d_{ij} is euclidian distance between i-th point of an image and j-point of a pattern. In order to find the nearest point in the pattern the minimum of values in the row is searched.

Stability criterion. Two points in the image and the pattern, which have the closest euclidian distance do not guarantee correct matching, that is why the new criterion was introduced. It checks stability. With the aim of verifying stability for a given characteristic point in an image, two closest characteristic points in the pattern are searched with distance values d_{ij} and d_{ik}.

$$d_{ij}\alpha < d_{ik} \tag{2}$$

If these points are too close (their distances satisfies the condition 2), it is difficult to choose the point the examined point corresponds to and this matching is rejected. The symbol α stands for stability coefficient (its value $\in [1.3; 1.8]$).

Neighborhood criterion. Its essence lies in defining maximal distance that can exist between the corresponding points. Even if the corresponding points satisfy the first and the second criteria their descriptors can differ much. Therefore the requirement that corresponding point should exist in neighbourhood of the examined point seems justifiable.

$$d_{ij} < \beta, \tag{3}$$

where β is the maximal distance between points of examined image and the pattern. For the feature vector with the length of 128 the value β is chosen from the range $[40000, 90000]$. However, after this phase the matched corresponding points still contain many incorrect joined pairs.

If the pattern exists in the image, with respect of geometry, it can be translated, rotated, scaled along one axis. The affine transform, should help to solve this problem. Practically, to define transformation we need to find 3 points satisfying the condition that each point $P = (x, y)$ from the image is mapped to the point $P^{'} = (x^{'}, y^{'})$ in the pattern. Once the transformation has been defined, the re-mapping proceeds by calculating the corresponding pattern coordinates. In many cases pixels can not be mapped precisely. To minimize this effect we introduced *imprecise affine transformation*. We assumed that the point from an image can be mapped in the neighborhood of the point $P^{'} = (x^{'}, y^{'})$. Because incorrect matching still has occurred, we have introduced 3 restrictions on affine transformation. Their interpretation using vector representation is shown in Fig. 3.

Restriction1– against excessive scaling along one axis. The condition expressed by (4) should be satisfied; d defines a vector length as presented in Fig. 3,

$$|\frac{d_1}{d_2} - \frac{d_3}{d_4}| < \gamma, \tag{4}$$

where γ is the maximal allowed difference in scaling. In practice $\gamma \in [0.05{,}0.5]$.

Restriction2 – against excessive slant transformation. It is performed by satisfying the condition 5, which limits the difference between angles;

$$|\lambda - \theta| < \epsilon \tag{5}$$

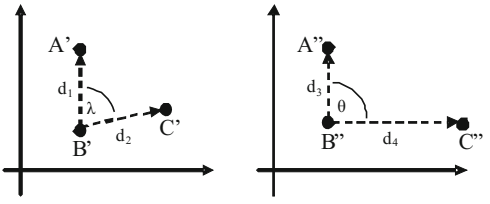


Fig. 3. Affine transformation presented in the vector form (points with ' are mapped to ")

λ and θ are the angles between vectors; d_1, d_2 and d_3, d_4 are the length of vectors as presented in Fig.3 ; ϵ is the assumed maximal difference between these angles. In practice $\epsilon \in [30°, 60°]$.

Restriction3 – against transformation instability. Once the angles λ and θ are close to 0° or 180°, the transformed points would not create linear relation because the imprecise affine transform is applied. That is why the possibility to achieve unstable transformation is eliminated by satisfying the conditions (6).

$$\begin{aligned} \lambda &> \phi \\ |\lambda - 180°| &> \phi \\ \theta &> \phi \\ |\theta - 180°| &> \phi \end{aligned} \quad . \tag{6}$$

The angles θ and λ have the same meaning as above,ϕ is the minimal angle between vectors, in our research $\phi \in [5°,20°]$.

Once the second phase is finished unification of affine corresponding sets is performed. The aim of this phase is to check which of the sets from the second phase corresponds to the same pattern in the image. The algorithm starts with the largest set which we have called the *core set*. The set will grow by joining other sets that have high enough number of common corresponding points. It is expressed by condition (7), where A and B are two corresponding points sets and ψ is a treshold of the common points number of the sets A and B.

$$|A \cap B| \geq \psi \tag{7}$$

The cores that have no common points assign pattern in the examined image.

2.2 Shape Information Processing

To acquire information about shapes in an image we have considered: Harris operator that produces location of corners and Hough transform that enables mathematical description of the searched pattern edges in the image.

Harris operator can produce many corners located very closely. To solve this problem clustering was applied. The clustering process is used as long as in the neighborhood r another corner exists. The coordinates of the new corner are the average of all clustered corners coordinates. A drawback of Harris operator is the fact that it does not produce any descriptor. There are methods that allow

Fig. 4. Three examples of the same object. On the left – information carried by color, in the middle - by shapes and shadows, on the right - by shape only.

to create such matching. They mainly examine the neighborhood of corners by calculating local gradient changes [4]. They are very efficient for images which include almost identical content.

Human being is able to recognize object on the basis of shapes. Let us consider an example presented in Fig. 4. Despite that from the left to the right each image contains less and less information we are still able to recognize an object. Perfect solution would be affine transform for corresponding corners. It would allow to combine information about color and shape. Because currently this problem has not been solved, we have reinforced information obtained from the SIFT algorithm by information about corners. In the proposed modification of the algorithm there is a step creating imprecise affine transform for each characteristic point. These transformations were used to check ability to match the pattern corners and image corners.

The second source of shape information comes from Hough transform which searches for lines in an image. They are represented by the slope and translation. On this basis the location of the beginning and the end of the line have to be found. The beginning is the point where density edge points exceeds the assumed treshold. It was set to 5 points. Then the points assigned in the image are compared to the analogical points of the pattern.

Similarly to the corners, the edges are compared on the basis of the SIFT imprecise affine transform but operating on unified corresponding sets. This change was applied to validate correct detection. For each affine transform assigning the pattern in the image the set of corresponding edges should be defined. Assuming that as an input we have the beginning and the end points of edges in the pattern and the image, the algorithm of edges comparison is as follows:

1.For each edge in the image map it to the pattern using affine transform
2. If at least one point of transformed edge is located in radius r from the point of pattern edge go to 3.; in other case go to 1.
3. If the angle σ is less than treshold, the compared edges are assigned as corresponding ones.
4. Add the lines from point 3. to the corresponding lines and go to the point 1.

SIFT algorithm, Harris operator and Hough transform extract features that correspond to the features in a pattern. Our idea was to map the image features onto the pattern features by the affine transform. To avoid some incorrect mapping in the transformation process the last phase – *validation* was introduced. Validation is made on he basis of relation between the number of features identified in the pattern and the number of features that have been joined with an image. This value is called the *level of covering*, which is calculated for each feature separately: SIFT descriptors, Harris operator and Hough transform. A pattern is identified if each of this covering levels is greater than the assumed tresholds. The value too high can cause that the method will not be able to identify patterns, while the value too small results in too many errors. The treshold values were assigned in the experimental study.

3 Experimental Study

The goal of experiments was to check efficiency of the method. In the initial experiment we performed tuning of the method parameters (tresholds values) and we checked their influence on the final results. In all experiments we used 50 patterns representing company logos and traffic signs. Together we have collected 100 images with 122 patterns. In each image at least one pattern has existed.

Treshold values. On the basis of experiments the tresholds for covering levels were set to the values presented in the Table 1. The highest value of covering level treshold was noticed for SIFT algorithm. This means that the algorithm offers the best selection in the context of correct identification. The tresholds for Harris operator and the Hough transform produce similar number of errors, but the last one has the greater number of correct identifications comparing to the former one. But still for all algorithms used in the experiments for one correct identification there exists incorrect one. To examine this problem the next experiment was done, in which we have checked the number of identified patterns in relation to treshold referred to the sum of all coverings (from all applied here algorithms). The results are shown in Fig 5. The vertical line assigns treshold value equal to 0.64. For this value the number of correct identification of patterns was 97 while incorrect one is equal to 76. In comparison to the result of SIFT algorithm, the proposed hybrid method gives much less incorrect classification (106-76=30). This result is 28% better than the result of classical SIFT algorithm. This comparison confirms that LLCPR is more effective than the single

Table 1. Tresholds for covering levels of algorithm applied in the method

algorithm name	treshold	correct identification	incorrect identification
SIFT	0.17	97	106
Harris operator	0.11	90	135
Hough transform	0.02	102	137

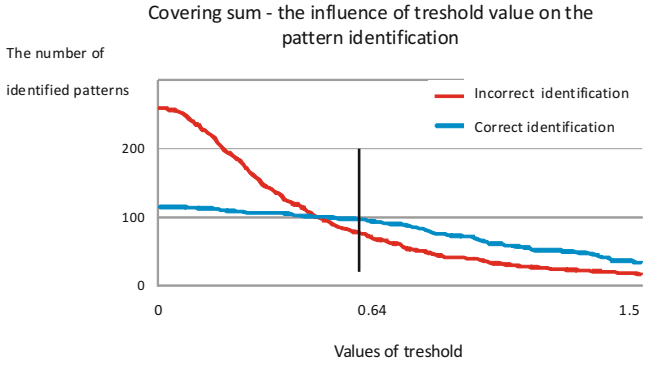


Fig. 5. Relationship between the number of identified patterns and treshold value for covering sum obtained on the basis of all applied algorithms

Table 2. Component measures for precision and accuracy calculation

Parameter	Meaning	Value
TP – true positive	correctly identified pattern	97
FP –false positive	incorrectly identified pattern	76
TN – true negative	correctly identified pattern absence	4827
FN – false negative	incorrectly detected absence of pattern	25

SIFT algorithm. Fig. 5 shows also that the number of correct classification is not greater for combined algorithm than in the modified SIFT algorithm. It is caused by the fact that affine transform for edges is created on the basis of the SIFT algorithm.

Efficiency of patterns identification. In this experiment precision and accuracy of the system were measured. Precision informs how much repeatable are the results while accuracy expresses the rate of correct predictions made by the model over a data set. Table 2 shows the results obtained in the experiment. Formally, accuracy can be expressed as is the proportion of true results (both true positives and true negatives) in the population:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

Precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives):

$$precision = \frac{TP}{TP + FP} \tag{9}$$

Fig. 6. The examples of correctly identified patterns

On the basis of the values from Table 2 accuracy is equal to 98,0% and precision 56,1%. The method has detected 79,5% of all patterns. When LLCPR analyses images that do not contain patterns, only 1,5% recognitions were wrong. Fig. 6 presents the results of identification made by LLCPR system. It is worth noticing that in spite of considerable differences between logos being the patterns and their representatives in the image the system successfully has found them in the images.

4 Summary

The basis of the system performance is the modified SIFT algorithm. It searches for imprecise affine transform that assigns location of a pattern in an image. The rest of applied algorithms helps to verify whether the LLCPR system has detected a pattern properly. To enlarge probability of pattern detection in the research, our idea was to develop a method of corners matching, but it has not been finished with success. For this reason we have reinforced information acquired by SIFT algorithm by information about corners and edges. It resulted in 28% reduction of incorrect detections number comparing to the classical SIFT algorithm. Such result is promising for further development of the method. However, to fully evaluate efficiency of the proposed approach it is necessary to extend the images data set.

Unluckily because of the lack of corner matching method, effectiveness of correct patterns identification was not increased. As a drawback of the method a high number of tresholds can be perceived, which should be adjusted. Their automatic assignment will be considered in the future. Also the future works will concentrate on improvement of the number of correct identified patterns. We consider to apply the new version of SIFT – ASIFT (Affine Scale Invariant Feature Transform), [8]. Another solution could be the application of perspective transform for the corner matching.

Acknowledgements. This work is partially financed from the Ministry of Science and Higher Education Republic of Poland resources in 2008 2010 years as a Poland-Singapore joint research project 65/N–SINGAPORE/ 2007/0.

References

1. Fang, C.Y., Chen, S.W., Fuh, C.S.: Automatic change detection of driving environments in a visionbased driver assistance system. IEEE Trans. Neural Networks 14(3), 646–657 (2003)
2. Franklin, S.: The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent. Society for Design and Process Science, CA (2006)
3. Furesjo F., Christensen H. I., Eklundh J. O., Object Recognition using Multiple Cues, CVAP - Computational Vision and Adaptive Perception, Stockholm (2004)
4. Haiyan, Y., Cuihua, R., Xiaolin, Q.: A New Corner Matching Algorithm Based on Gradient. School of Electronics and Information Technology. Harbin Institute of Technology, Harbin (2008)
5. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. In: Touretzky, D. (ed.) Advances in Neural Information Processing Systems 2, pp. 396–404. Morgan Kaufman, Denver (1990)
6. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
7. Lowe, D.G.: Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, U.S. Patent 6,711, 293 for the SIFT algorithm
8. Morel, J.M., Yu, G.: ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences 2(2) (2009)
9. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: IEEE CVPR, vol. 1, pp. 511–518 (2001)
10. Wang Y.: The Cognitive Informatics Theory and Mathematical Models of Visual, Information Processing in the Brain. Cognitive Informatics and Natural Intelligence (2009)

Rule Induction Based-On Coevolutionary Algorithms for Image Annotation

Paweł B. Myszkowski

Wrocław University of Techonology,
Wyb.Wyspiańskiego 27, Wrocław, Poland
pawel.myszkowski@pwr.wroc.pl
http://www.ii.pwr.wroc.pl/~myszkows/

Abstract. This paper describes our experiments in the field of evolutionary algorithms for rule extraction applied to automating image annotation and classification problems. Presented approach is based on classical evolutionary algorithm with binary representation of 'if-then' rules. We want to show that some search space reduction techniques make possible to get problem's solution. Paper shows that the gap between classification and image annotation problem can be bridged easily. Some experiments with EA parametrization in image annotation problem are presented. There are presented first results on ECCV 2002 dataset in image annotation problem.

Keywords: data mining, evolutionary algorithms, rule induction, classification, image annotation.

1 Introduction

The size of datasets is growing constantly and cannot be analyzed in simple way, so we use automating process, so-called Knowledge Discovery in Databases (KDD) useful tool in Machine Learning domain. The most interesting, from this paper point of view, is its one stage data mining (DM). The data mining is an interdisciplinary field and its essence is knowledge acquisition from large amount of data. As our data might contain useful hidden and implicit knowledge, the extracted knowledge can be successfully used in very important real-world domains such as image annotation problem. The image annotation problem [1] [3] [5] [10] is process of images labeling that uses pre-defined keywords (usually represent semantic content). The process is laborious and error prone thus it is strongly recommended to build automated classification (labeling) mechanisms. In literature Evolutionary Algorithms (EA) are used to feature selection in image annotation problem [12] but according to our best knowledge in the literature there is no EA application to the image annotation problem by rule induction.

EA are metaheuristics that explore the solution search space and can be successfully applied for data mining tasks, such as clustering, prediction, classification and rule induction (work [7] contains a survey of EA applications to KDD). Rule discovering (or rule induction) is one of the most studied data mining task

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 232–241, 2011.

and the main goal is to build model of given data that describes dataset with the best possible accuracy. Such model can be based on an intuitive 'if-then' rules where: if-part (antecedent) contains attribute conditions and then-part (consequent) that contains the predicted class value (label). The classifier quality (accuracy of the gained model) can be tested on unseen data and measured by the prediction error value. EA is widely used in data mining tasks, e.g. [2] [3] [4] [17].

Basically, EA codes the problem solution as an individual and operates its features using genetic operators (defined usually as mutation and crossover). The quality of individual is given as fitness function formula and its better value gives the higher probability of getting an offspring. In the rule discovering task the main motivation is discovering the rules with the high predictive accuracy value. In literature we can find some approaches based on natural evolution that extend simple EA (e.g. classical genetic algorithm [8] [13]) by some additional elements. For instance, commonly used *if-then* form of rule can be defined a single individual (so-called Michigan approach) or included in ruleset form of individual (Pittsburgh approach). Mostly, there is used the Pittsburgh approach as it takes into consideration rule's interactions, also is much more natural and intuitive.

Interesting EA based method, so-called Genetic Programming can be found in [3], where individual is represented by logic tree that corresponds to logical expression such as rule:

$$If\ attrib_0 > 5\ and\ (attrib_3 = 2,5\ or\ atrrib_2 < 1,3)\ then\ class_3 \qquad (1)$$

to describe attributes' conditions in given class. The rule is presented as decision tree, where the internal nodes become operators and each leaf node represents attributes and corresponding condition values.

The multipopulation EA is based on the natural phenomena of coevolution (Coevolutionary Algorithm), where evaluation of few populations runs effectively in parallel way. The main motivation in such approach is EA computational cost reduction or/and evolution by problem decomposition. The method presented in [17] operates on populations that correspond to ruleset of n rules (where n is a parameter) linked by token competition as a form of the niching method. Paper [15] describes approach that takes into consideration distributed genetic algorithm for rule extraction task and it is proved a positive influence of dynamic data partitioning distribution model to final classifier accuracy.

There is also another strong trend in EA in DM: specialized genetic operators usage. In Pittsburgh model in classification task it is very important that individual representation consists of complete classifier where rules cooperate in the whole ruleset and the genetic operator causing its separation may make worse its classification accuracy. In [17] this problem is discussed, there is given a proposition of symbiotic combination operator which is kind of heuristic that analyzes results of changes in newly created individual. Another type of genetic operator specialization can be usage of some hill-climbing algorithms to improve individuals (as candidate solution) by making minor modifications: if it causes fitter individual, given change is accepted. However, this causes the Baldwin

effect [12]. In the evolution process ruleset can be modified in pruning procedure as well. For instance, in work [5] is optimized by removing unused/invalid attribute condition according to information gain measure value by examination of some small changes results in the ruleset. Also, we can find hybridization as a quite strong trend, where the main motivation of such propositions is to build approach that links advantages of connected methods.

The remainder of this paper is organized as follows. Details of problem definition and our approach for rule discovering task is presented in section 2. Research methodology, used benchmark datasets and results of experiments are presented in section 3. Finally, section 4 presents conclusions and future research directions.

2 CAREX: Coevolutionary Algorithm for Rule EXtraction

Our proposed approach uses standard EA schema and starts the learning process with initial population (usually created randomly), in the next step individuals of current population are evaluated: each individual receives a fitness function value according to quality of containing solution. Next, EA checks if stop conditions are not met: usually it is limit of generations or the best individual fitness value is acceptable (success). If stop criteria is not met EA runs the selection procedure that gives a seed of the new generation; then it is provided a communication between individuals (by crossover operator) and works the independent trial (by mutation operator). The whole process repeats until some stopping condition is met. The crucial issue in evolutionary based method is definition of individual representation schema, genetic operators (mutation and crossover) and evaluation function form, that gives information about the individual fitness function value. In this section above elements are described.

2.1 Representation Schema

In CAREX approach we decided to construct individual representation schema as simple as possible to get reduction of solution search space size. This is gained by coding of arguments value in binary representation (usually 8 bits codes value) and only two opposite logical operators: IN and NOT_IN (uses only one bit). Such methodology makes possible to use simple EA and what we wanted to show is that there is no need of EA extension to get acceptable solution. Also we wanted to examine if there is strong need of genetic operators specialization in data mining applications.

In our research we use only the Pittsburgh model to get all rules interactions in one individual. Therefore individual is represented as set of rules (ruleset) that can assign instance to one class or gives model of all classes presented in dataset and consists of rules connected to all class identifiers as follows:

$$RuleSet := \{Rule_0, \ldots, Rule_n\} \tag{2}$$

Each rule is represented as set of commonly used if-then type rules as follows:

$$Rule_i := IF\ A_0\ and \ldots and\ A_n\ THEN\ class_j \tag{3}$$

where $class_j$ symbol represents given class identifier, and $A_i,\ i \in \{0, \ldots, n\}$ represents numeric range for a condition for *i-th* attribute as follows:

$$A_i := attribute_i\ operator\,(a, b) \tag{4}$$

where bellow operator can be:

$$IN\ \ (a, b) \rightarrow < attr_i < b\ \ or\ \ \ NOT\_IN\,(a, b) \rightarrow attr_i < a\ or\ attr_i > b \tag{5}$$

where $a < b\ and\ a, b \in \Re$. Above representation is strictly based on conditions combination for selected attributes. We decided to use only two operators to keep individual representation as simple as possible. For instance, rule is:

$$IF\ attr_0\ IN\,(0.1, 0.5)\ \ and\ attr_2\ NOT\_IN\ \ (1.0, 1.2)\ \ THEN\ class_2 \tag{6}$$

Above rule describes all instances with values from range $< 0,1; 0,5 >$for attribute 0. Another condition takes into consideration attribute 2, where its value cannot be in range $< 1,0; 1,2 >$ Data described by conjunction of conditions are proposed to label with $class_2$.

In CAREX we decided to use binary vector representation thus we are allowed to use classical binary genetic operators to manipulate individual's particles. To avoid a drastic change of attribute value we use a Gray code. Also each attribute is extended by one enabled/disabled bit. Before CAREX starts, the dataset is preprocessed: instances are analyzed to recognize domains for all attributes. Then, each attribute domain is mapped into binary vector. That allows to keep each individual valid and there is no need to waste extra CPU time for repairing or removing invalid attribute values.

In proposed representation we use only AND logical operator, therefore EA to describe some set of instances as two separate conditions use ruleset as two connected rules. Indeed, the relation between these rules is logical disjunction, and indeed there exist sort of rules coevolution phenomena.

2.2 Fitness Function

Generally, in EA, fitness function evaluation is very critical issue. Its definition decides about shape of solution landscape and must be defined very carefully. As rule extraction problem corresponds to data mining and our individual is a ruleset we use commonly used **classification** measure. In classification the main goal is prediction of the value $c_i \in C$ (class) analyzing values of attributes $x_i$of given instance $x_i = \{x_i^0, \ldots, x_i^n\}$ where $x_n \in X$ defines solution landscape. Thus classification task is based on explore set of $(x_1, c_1), \ldots, (x_n, c_m)$ to build model $m\,(x_i) : X \rightarrow C$ that labels unseen instance $x_k \in X$. Evaluation of rule is connected to its quality as classifier. In such context of data mining domain,

the terms true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) we can define *recall* :

$$recall = \frac{tp}{tp + fp} \tag{7}$$

and *precison* measure as follows:

$$precision = \frac{tp}{tp + fn} \tag{8}$$

The recall value tells only if given rule labels instances properly, precision informs if rule covers all labeled data by proper class identifier. Above formulas say a lot about rule classification quality but there are two separate values. Although EA should use only one value to evaluate rule in literature (e.g. [3]) we can find some combinations of these values. However, we decided to use measure based on modified van Rijsbergen's effectiveness measure [16] (Fsc), than can be used in data mining too because it combines *precision* and *recall*:

$$Fsc = \frac{1}{\frac{\alpha}{precision} + \frac{1-\alpha}{recall}} \tag{9}$$

where α can give a predominance of one of two elements, but we established its value on 0,5 to keep two elements equal. If Fsc value is near 1 it means that evaluated rule has high quality as it corresponds to maximizing problem. The Fsc measure is very useful as fitness function form, but for comparison of gained results in literature is used other measure of predictive accuracy, as a rule quality measure, defined as follows:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{10}$$

where *tp/tn/fp/fn* corresponds respectively to true positive/true negative/false positive/false negative to denote accuracy of classification in given set of instances. Our experiments showed that Acc usage as fitness function formula is less effective than Fsc. However, in such form of fitness function occurs some distortions specially when dataset is dominated by one class and others have small representations.

Automating **image annotation** problem is strongly connected to classification problem but the main difference is that given image (as an instance to be annotated) may be labeled by more than one class (keyword). Also, image consists of set of parts (so-called segments) that may be labeled independently. Our approach in rule induction process is based on image segment annotation and makes possible to label each segment without any keyword domination, as it exists in classical learning process based on image. Also gained rules describe each keyword by attributes value condition and are easy to understand (in opposite to large decision tree), can be further analyzed and processed by human/expert systems. In the image annotation problem Fsc is used as solution quality measure and we use it as fitness function.

2.3 Genetic Operators

Our individual is represented by binary vector which can be operated by classic binary operators in simple way. Random change of selected bit works as mutation and inserts new information into chromosome. To deliver a communication in the CAREX population we developed uniform crossover (UX) swaps the parential's parts with the same probability, random based One-Cut Crossover (OX) that links two individuals to build the new one as combination of their genes. As all individuals have the same size, there is no situation where invalid individual is created by the random cut position.

3 Computational Experiments and Results

Evaluation of learning method is important to compare results against other methods. This can be done experimentally so we developed in Java a research environment that supports learning and rule validating process. First, there are used train data to generate *RuleSet* by CAREX to get possible high accuracy (*Fsc* is used as fitness function). Learning process based on evolution runs using selection, mutation and crossover operators. For evaluation accuracy of gained rules there is used train dataset, but when evolution process is finished the test dataset is used to validate predictive accuracy of generated *RuleSet*.

As CAREX in first stage was tested on benchmark UCI ML Repository[1] datasets (see [14]) we decided to define extra series of experiments to get more information about the CAREX effectivness. Artificially generated monks datasets [18] give us a great opportunity to make the reference to the other learning algorithms. Moreover, our special attention is to investigate how the standard genetic operators and parameters influence to presented method results. To examine of CAREX effectiveness in image annotation problem we use benchmark ECCV 2002 dataset [6].

3.1 Classification Problem - Monks Datasets

The three monks datasets [18] were created to compare learning algorithms in classification task. The problem involves a binary function defined over this domain. Each dataset is defined by 6 attributes and consists of train and test data (432 instances randomly selected) and defines other function. The monks1 dataset is the simple one and can be solved by 4 rules. The second dataset is more difficult, as generated 15 rules include some interactions between attributes. The last dataset (monks3) consists of noisy data (5%), and it may cause solving difficulties. The CAREX results gained on monks datasets are presented in Table 1.

Results gained by CAREX on monks datasets are satisfactory and shows high accuracy of presented metod and can be successfully compared to other classification methods. Datasets monks1 and monks3 are solved completely if we accept

[1] UCI Machine Learning Repository: http://www.ics.uci.edu/~mlearn/

Table 1. The best results of CAREX

data	avg.accuracy	CAREX configuration
monks1	100% ±0,0	rules=4 pop_size=10 gener.=100 000 Pux=0,3 Pm=0,01
monks2	86,16% ±1,6	rules=15 pop_size=100 gener.=10 000 Pux=0,3 Pm=0,01
monks3	97,22% ±0,0	rules=4 pop_size=5 gener.=10 000 Pux=0,3 Pm=0,01

that generated rules cannot manage well with noisy data. The second dataset gives a clue that a larger ruleset requires longer evolution process connected to larger population. CAREX work on monks2 dataset give results comparable to other learning methods, where the worst methods achieve accuracy equal to 57,2% and the best one achieves 100% [8].

3.2 Automating Image Annotation Problem - ECCV 2002

The ECCV 2002 dataset [6] consists of 4500 images (extra 500 images are given in test dataset) where each of them is described by 3-10 segments and each segment includes 36 attributes. The training dataset uses 263 keywords and each image is annotated by 3 - 4 keywords.

We applied CAREX to ECCV 2002 using keywords decomposition where the goal is to describe given keywords as attributes value conditions connected to rules. We analysed first results of 5 most frequent keywords to get method's accuracy and compare results to benchmark CRM [10] image annotation method. The CAREX results are given in Table 2, where data are given for test dataset accuracy and were repeated 10 times. It is worth to mention that average *Fsc* value in presented methods are comparable, moreover some CAREX configurations cause that presented method gives average better solutions than CRM.

However, still open issue is CAREX parametrization. We use only three basic parameters (number of rules, number of individuals and generation number) and three less problematic (mutation ratio, crossover ratio and tournament size). As the second group of parameters can be established experimentally using other (simpler) datasets, the first group needs experiments with ECCV 2002 that are very computational time consuming. Evolution process character (see Fig.1) suggests that problem landscape is searched effectively moreover in this stage of research we cannot find an optimal configuration[2] for all keywords (see Table 2).

The evolution process character (presented on Fig. 2) shows that evolution uses the whole time for experiments to get better solution, however the higher value of *Fsc* gained on train dataset not always gives higher value of *Fsc* on test dataset. Even more, time (generations) for EA cannot avoid this effect that can compared to overlearning phenomena well known in artificial neural networks. It is also partially confirmed by a relative high value of the standard deviation of *Fsc* value.

[2] CAREX configuration is given as follows: number_of_rules x numer_of_individuals x generations. In all used configurations mutation ratio Pm=0.01 and crossover ratio Pux=0,3.

Table 2. The averaged CAREX results gained on 5 most frequent keyword of benchmark ECCV 2002 image dataset

keyword	CAREX configuration 1x10x100	3x10x100	5x5x1000	5x10x1000	5x20x200	5x100x100	CRM
water	42,83% ±1,1	42,77% ±2,17	41,99% ±1,43	42,71% ±1,4	43,05% ±0,9	**43,75%** ±1,9	44,05%
sky	46,3% ±2,29	43,14% ±1,6	43,15% ±1,3	**48,29%** ±2,21	47,32% ±1,7	47,76% ±2,8	53,75%
tree	33,14% ±1,0	33,21% ±1,95	35,84% ±1,22	**43,64%** ±2,4	32,07% ±2,09	34,57% ±3,11	36,30%
people	36,0% ±2,6	35,02% ±2,2	34,67% ±1,94	38,72% ±2,0	**39,62%** ±2,7	37,83% ±3,2	42,10%
buildings	27,86% ±3,4	29,17% ±2,59	34,90% ±1,78	**38,66%** ±0,3	37,36% ±2,5	31,5% ±5,97	31,49%
average	37,23%	36,66%	38,10%	42,4%	39,99%	39,04%	41,54%

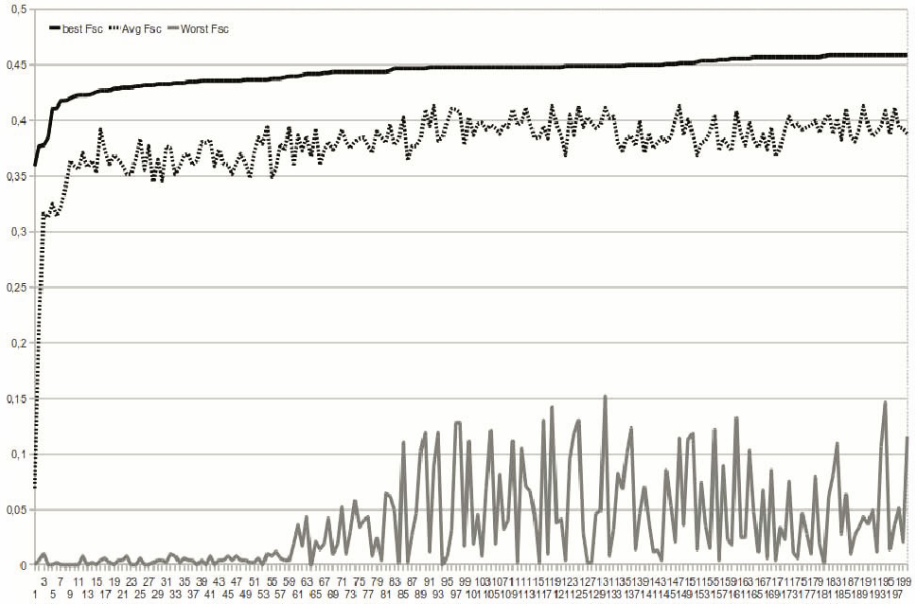


Fig. 1. The CAREX evolution character in ECCV (5rules, pop_size=100, gener.=200)

Although some parametrization EA problems, we find presented results as very promising as CAREX outperforms benchmark CRM results. We can see a great potential of presented method and this encourage us to examine method in the more complex way using all ECCV keywords. Other image annotation datasets experiments are strongly needed.

4 Summary and Further Work

The paper describes results of first experiments of EA application to image annotation problem and method has great potential in this area. Gained results in ruleset form are easy to analyze and understand also it can be used in further work as knowledge automatically generated by data mining in expert systems or human being. Further work consists of complex CAREX experiments based on ECCV 2002 dataset. In the next research stage other image annotation datasets will be used to examine CAREX effectivess in empirical way.

CAREX approach is based on EA and all parametrization EA problems occur there. Also EA as searching problem landscape method, is time consuming it gives the main directions for further work. Computation time should be reduced by EA techniques (e.g. local search methods and/or hybridization) but also by low level programming techniques (such as individuals caching or memoisation). We are working on GPU architecture implementation in CAREX to make the computation parallel and less time consuming.

Acknowledgments. This work is partially financed form the Ministry of Science and Higher Education Republic of Poland resources in 2008-2010 years as a Poland-Singapore joint research project 65/N-SINGAPORE/2007/0.

References

1. Alham, N.K., Li, M., Hammoud, S., Qi, H.: Evaluating Machine Learning Techniques for Automatic Image Annotations. In: Proc. of the 6th Inter. Conf. on Fuzzy Syst. and Knowledge Discovery, vol. 07, pp. 245–249 (2009)
2. Ang, J.H., Tan, K.C., Mamun, A.A.: An evolutionary memetic algorithm for rule extraction. Expert Systems with Applications 37, 1302–1315 (2010)
3. Bojarczuk, C., Lopes, H., Freitas, A.: Genetic programming for knowledge discovery in chest pain diagnosis. IEEE Eng. Med. Mag. 19(4), 38–44 (2000)
4. Cattral, R., Oppacher, F., Graham Lee, K.J.: Techniques for Evolutionary Rule Discovery in Data Mining. IEEE Congress on Evolution. Comp., Norway (2009)
5. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(3), 394–410 (2007)
6. ECCV 2002 image set, http://kobus.ca/research/data/eccv2002/
7. Freitas, A.A.: A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. In: Advantages in Evolutionary Computing: Theory and Aplications, pp. 819–845. Spinger, NY (2003)
8. Goldberg, D.: Genetic algorithms in search, optimization and machine learning. Addision Wesley, London (1989)
9. Halavati, R., Souraki, S.B., Esfandiar, P., Lotfi, S.: Rule Based Classifier Generation using Symbiotic Evolutionary Algorithm. In: 19th IEEE Inter. Conf. on Tools with AI. ICTAI, vol. 1, pp. 458–464 (2007)
10. Lavrenko, V., Manmatha, R., Jeon, J.: A Model for Learning the Semantics of Pictures. In: Proceedings of Advance in Neutral Information Processing (2003)

11. Liu, J.J., Kwok, J.T.: An extended genetic rule induction algorithm. In: Proc. of the Congress on Evol. Comp. (CEC), La Jolla, California, USA, pp. 458–463 (2000)
12. Lu, J., Zhao, T., Zhang, Y.: Feature selection based-on genetic algorithm for image annotation. Knowledge-Based Systems 21, 887–891 (2008)
13. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Heidelberg (1994)
14. Myszkowski, P.B.: Coevolutionary Algorithm for Rule Induction. In: Proc. of the 5th Inter. Symp. Advances in Artificial Intelligence and App., AAIA 2010, pp. 73–79. IEEE, Los Alamitos (2010)
15. Rodriguez, M., Escalante, D.M., Peregrin, A.: Efficient Distributed Genetic Algorithm for Rule Extraction. Applied Soft Computing 11(1) (2011)
16. van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth, London (1979)
17. Tan, K.C., Yu, Q., Ang, J.H.: A coevolutionary algorithm for rules discovery in data mining. Inter. Jour. of Systems Science 37(12), 835–864 (2006)
18. Thrun S.B., et al.: The MONK's Problems - A Performance Comparison of Different Learning algorithms, Technical Report CS-CMU-91-197, Carnegie Mellon Univ. (1991)

Complex Fuzzy Computing to Time Series Prediction — A Multi-Swarm PSO Learning Approach

Chunshien Li and Tai-Wei Chiang

Laboratory of Intelligent Systems and Applications
Department of Information Management
National Central University, Taiwan, ROC
jamesli@mgt.ncu.edu.tw

Abstract. A new complex fuzzy computing paradigm using complex fuzzy sets (CFSs) to the problem of time series forecasting is proposed in this study. Distinctive from traditional type-1 fuzzy set, the membership for elements belong to a CFS is characterized in the unit disc of the complex plane. Based on the property of complex-valued membership, CFSs can be used to design a neural fuzzy system so that the system can have excellent adaptive ability. The proposed system is called the complex neuro-fuzzy system (CNFS). To update the free parameters of the CNFS, we devise a novel hybrid HMSPSO-RLSE learning method. The HMSPSO is a multi-swarm-based optimization method, first devised by us, and it is used to adjust the premise parameters of the CNFS. The RLSE is used to update the consequent parameters. Two examples for time series foresting are used to test the proposed approach. Through the experimental results, excellent performance has been exposed.

Keywords: complex fuzzy set (CFS), complex neuro-fuzzy system (CNFS), hierarchical multi-swarm particle swarm optimization (HMSPSO), recursive least square estimator (RLSE), time series forecasting.

1 Introduction

A time series is a sequence of historical statistical observations, for example, oil prices and stock prices in financial market. Time series analysis is to estimate the statistical regularities existed within the observed data and their connection to future tendency. The purpose of time series model is to explore a possible functional relationship with which the historical data are connected to future trend, so that a decision-making can be made in advance. Because accurate forecasting for the future trend is usually difficult in complex and nonlinear real-world problems, many researchers have used intelligent computing methods for time series forecasting, where fuzzy theory and neural networks have been widely investigated [1]-[4]. Although neural networks have excellent mapping ability and link-type distributed structure, they are usually considered as black-box systems, which are not easy to explain with human's knowledge. In contrast, fuzzy inference systems, providing a complementary alternative to neural networks, can extract human's experience and knowledge to form If-Then rules, which are easy to be explained by human. Both neural network and fuzzy system are with the property of universal approximator. Consequently, they

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 242–251, 2011.

can be integrated as a neuro-fuzzy system (NFS) [1], which incorporates the advantages of fuzzy inference and neural-learning. The theory of NFS has become popular and important to modeling problems [1], [3], [5].

In this study, we propose a novel complex neuro-fuzzy system (CNFS) using the theory of complex fuzzy set (CFS) to achieve high prediction accuracy for the problem of time series forecasting. The concept of CFS is an extension from the standard type-1 fuzzy set whose membership is in the real-valued interval of [0, 1]. The membership for elements belong to a complex fuzzy set is expanded to the complex-valued unit disc of the complex plan [6]. This property can be used to augment the adaptability of the proposed CNFS, if compared to its counterpart in NFS form. Although the CFS theory has been proposed [6]-[9], it is hard to construct intuitively understandable complex fuzzy sets for application. The theoretical curiosity on CFS remains. For this reason, we propose the CNFS using the theory of CFS to study its adaptability gain for the problem of time series forecasting. For the training of the proposed CNFS, we devise a new learning method, which combines the novel hierarchical multi-swarm particle swarm optimization (HMSPSO) algorithm with the recursive least square estimator (RLSE) algorithm. The HMSPSO method is devised by us and presented in this paper. It is used to update the premise parameters of the proposed CNFS. In the meanwhile, the RLSE is used to adjust the consequent parameters. The HMSPSO is different from another multiple-swarm version of PSO in [10]. The HMSPSO is devised in hierarchical form to enhance searching multiplicity and to improve the drawback of the standard PSO. With the hybrid HMSPSO-RLSE learning method, it can effectively find the optimum solution for the parameters of the CNFS. Two time series examples are used to test the prediction performance by the proposed approach. The proposed approach shows not only better adaptability in prediction performance than its traditional NFS counterpart, but also the superiority to other compared approaches [11].

The paper is organized as follows. In Section 2, the proposed complex neuro-fuzzy using complex fuzzy sets is specified. In Section 3, the HMSPSO-RLSE hybrid learning method is given. In Section 4, two examples for time series forecasting are given. Finally, the paper is discussed and concluded.

2 Methodology for Complex Neuro-Fuzzy System

The proposed complex neuro-fuzzy system (CNFS) using complex fuzzy sets is in inheritance of the benefits of both fuzzy inference system and neural network, especially the ability of being universal approximator that can approximate any function to any accuracy theoretically [12]-[13]. For a CFS, the membership state for elements belong to the CFS is within the complex-valued unit disc of the complex plane. This is in contrast with a traditional type-1 fuzzy set, to which the membership for elements belong is within the real-valued unit interval [0, 1]. In the following, we first introduce the notation of CFS, and then present the theory of the proposed CNFS.

2.1 Complex Fuzzy Set

The theory of complex fuzzy set (CFS) can provide a new development for fuzzy system research and application [6]-[9]. The membership of a CFS is complex-valued,

different from fuzzy complex numbers developed in [14] . The membership function to characterize a CFS consists of an amplitude function and a phase function. In other words, the membership of a CFS is in the two-dimensional complex-valued unit disc space, instead of in the one-dimensional real-valued unit interval space. Thus, CFS can be much richer in membership description than traditional fuzzy set. Assume there is a complex fuzzy set S whose membership function $\mu_s(h)$ is given as follows.

$$\begin{aligned}\mu_s(h) &= r_s(h)\exp\left(j\omega_s(h)\right) \\ &= \mathrm{Re}(\mu_s(h)) + j\mathrm{Im}(\mu_s(h)) \\ &= r_s(h)\cos\left(\omega_s(h)\right) + jr_s(h)\sin\left(\omega_s(h)\right)\end{aligned} \tag{1}$$

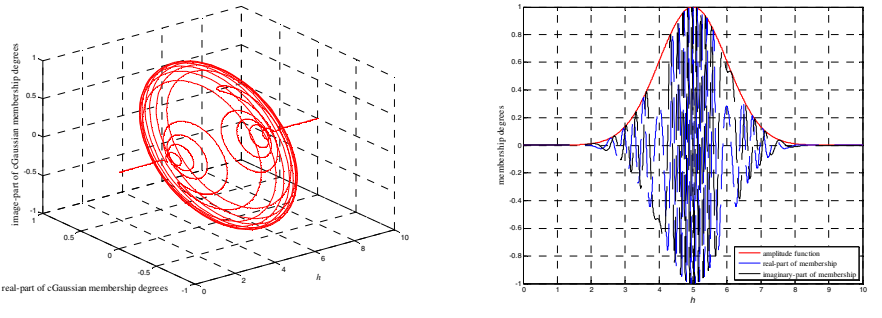
where $j = \sqrt{-1}$, h is the base variable for the complex fuzzy set, $r_s(h)$ is the amplitude function of the complex membership, $\omega_s(h)$ is the phase function. The property of sinusoidal waves appears obviously in the definition of complex fuzzy set. In the case that $\omega_s(h)$ equals to 0, a traditional fuzzy set is regarded as a special case of a complex fuzzy set. A novel Gaussian-type complex fuzzy is designed in the paper, and an illustration for a Gaussian-type complex fuzzy set is shown in Fig. 1. The Gaussian-type complex fuzzy set, denoted as $cGaussian(h, m, \sigma, \lambda)$, is designed as follows.

$$cGaussian(h,m,\sigma,\lambda) = r_s(h,m,\sigma)\exp(jw_s(h,m,\sigma,\lambda)) \tag{2a}$$

$$r_s(h,m,\sigma) = Gaussian(h,m,\sigma) = \exp\left[-0.5\left(\frac{h-m}{\sigma}\right)^2\right] \tag{2b}$$

$$w_s(h,m,\sigma,\lambda) = -\exp\left[-0.5\left(\frac{h-m}{\sigma}\right)^2\right] \times \left(\frac{h-m}{\sigma^2}\right) \times \lambda \tag{2c}$$

In (2a) to (2c), h is the base variable and $\{m, \sigma, \lambda\}$ are the parameters of mean, spread and phase frequency factor for the complex fuzzy set.



(a) (b)

Fig. 1. Illustration of Gaussian-type complex fuzzy set. (a) 3-D view with the coordinates of base variable, real-part membership and imaginary-part membership. (b) Amplitude membership and imaginary-part membership vs. base variable.

2.2 Theory of Complex Neuro-Fuzzy System

Suppose an *M*-input-one-output complex fuzzy system is designed with *K* first-order Takagi-Sugeno (T-S) fuzzy rules, given as follows.

$$\begin{aligned} &\text{Rule } i\text{: IF } (x_1 \text{ is } A_1^i(h_1)) \text{ and } (x_2 \text{ is } A_2^i(h_2)) \ldots \text{and } (x_M \text{ is } A_M^i(h_M)) \\ &\quad \text{Then } z^i = a_0^i + \sum_{j=1}^{M} a_j^i h_j \end{aligned} \tag{3}$$

for i =1,2,...,K, where x_j is the j-th input linguistic variable, h_j is the j-th input base variables, $A^i_j(h_j)$ is the complex fuzzy set for the j-th premise condition in the i-th rule, z^i is the output of the i-th rule, and $\{a^i_j, i=1,2,\ldots,K$ and $j=0,1,\ldots,M\}$ are the consequent parameters. The fuzzy inference of the complex fuzzy system can be cast into neural net structure with six layers to become the complex neuro-fuzzy system (CNFS). The explanation for the six layers is specified as follows.

Layer 1: The layer is called the input layer, which receives the inputs and transmits them to the next layer directly. The input vector at time t is given as follows.

$$\mathbf{H}(t)=[h_1(t), h_2(t), \ldots, h_M(t)]^{\mathrm{T}} \tag{4}$$

Layer 2: The layer is called the fuzzy-set layer. Each node of layer represents a linguistic value characterized by a complex fuzzy set for the premise part of the CNFS and to calculate the complex membership degrees $\{\mu^i_j(h_j(t)),\ i=1,2,\ldots,K$ and $j=0,1,\ldots,M\}$. The complex fuzzy sets $A^i_j(h_j)$ can be designed using the Gaussian-type of complex membership function given in (2a) to (2c).

Layer 3: This layer is for the firing-strengths of fuzzy rules. The nodes perform the *fuzzy-product* operations for the firing strengths of the fuzzy rules. The firing strength of the i-th rule is calculated as follows.

$$\begin{aligned} \beta^i(t) &= \mu_1^i\big(h_1(t)\big) * \mu_2^i\big(h_2(t)\big) * \ldots * \mu_{\mathrm{M}}^i(h_M(t)) \\ &= \prod_{j=1}^{M} r_j^i \left(h_j(t)\right) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right) \end{aligned} \tag{5}$$

where r^i_j is the amplitude of complex membership degree for the j-th fuzzy set of the i-th rule. With (5), the $\omega_{A_1^i \cap \ldots \cap A_M^i}$ is calculated.

Layer 4: This layer is for the normalization of the firing strengths. The normalized firing strength for the i-th rule is written as follows.

$$\lambda^i(t) = \frac{\beta^i(t)}{\sum_{i=1}^{K} \beta^i(t)} = \frac{\prod_{j=1}^{M} r_j^i \left(h_j(t)\right) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{i=1}^{K} \prod_{j=1}^{M} r_j^i \left(h_j(t)\right) \exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \tag{6}$$

Layer 5: The layer is called the consequent layer for calculating normalized consequents. The normalized consequent of the i-th rule is given as follows.

$$\begin{aligned} \xi^i(t) &= \lambda^i(t) \times z^i(t) \\ &= \lambda^i(t) \times \left(a_0^i + \sum_{j=1}^{M} a_j^i\, h_j(t)\right) \\ &= \frac{\prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{i=1}^{K} \prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \times \left(a_0^i + \sum_{j=1}^{M} a_j^i\, h_j(t)\right) \end{aligned} \tag{7}$$

Layer 6: This layer is called the output layer. The normalized consequents from Layer 5 are congregated into the layer to produce the CNFS output, given as follows.

$$\begin{aligned}\xi(t) &= \sum_{i=1}^{K} \xi^i(t) = \sum_{i=1}^{K} \lambda^i(t) \times z^i(t) \\ &= \sum_{i=1}^{K} \frac{\prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)}{\sum_{i=1}^{K}\prod_{j=1}^{M} r_j^i\left(h_j(t)\right)\exp\left(j\omega_{A_1^i \cap \ldots \cap A_M^i}\right)} \times \left(a_0^i + \sum_{j=1}^{M} a_j^i\, h_j(t)\right)\end{aligned} \tag{8}$$

Generally the output of the CNFS is complex-valued and can be expressed as follows.

$$\begin{aligned}\xi(t) &= \xi_{\mathrm{Re}}(t) + j\xi_{\mathrm{Im}}(t) \\ &= |\xi(t)| \times \exp(j\omega_\xi) \\ &= |\xi(t)| \times \cos\left(j\omega_\xi\right) + j|\xi(t)| \times \sin(j\omega_\xi)\end{aligned} \tag{9}$$

where $\xi_{\mathrm{Re}}(t)$ is the real part of the output of the CNFS, and $\xi_{\mathrm{Im}}(t)$ is the imaginary part. Based on (9), the complex inference system can be viewed as a complex function, expressed as follows.

$$\xi(t) = \mathrm{F}(\mathbf{H}(t), \mathbf{W}) = \mathrm{F}_{\mathrm{Re}}(\mathbf{H}(t), \mathbf{W}) + j\mathrm{F}_{\mathrm{Im}}(\mathbf{H}(t), \mathbf{W}) \tag{10}$$

where $\mathrm{F}_{\mathrm{Re}}(.)$ is the real part of the CNFS output, $\mathrm{F}_{\mathrm{Im}}(.)$ is the imaginary part of the output, $\mathbf{H}(t)$ is the input vector to the CNFS, $\mathbf{W}$ denotes the parameter set of the CNFS, which is composed of the subset of the premise parameters and the subset of the consequent parameters, denoted as $\mathbf{W}_{\mathrm{If}}$ and $\mathbf{W}_{\mathrm{Then}}$, respectively.

$$\mathbf{W} = \mathbf{W}_{\mathrm{If}} \cup \mathbf{W}_{\mathrm{Then}} \tag{11}$$

3 Multi-Swarm-Based Hybrid Learning for the Proposed CNFS

We devise a hybrid learning method including a multi-swarm-based particle swarm optimization and the recursive least squares estimator (RLSE) method to update the $\mathbf{W}_{\mathrm{If}}$ and $\mathbf{W}_{\mathrm{Then}}$, respectively. Particle swarm optimization (PSO) is a swarm-based optimization method [15]-[18], motivated by the food searching behavior of bird flocking. There are many particles in a PSO swarm. The best location for a particle in the search process is denoted as **pbest**. The particles in the swarm compete each other to become the best particle of the swarm, whose location is denoted as **gbest**. In this study, we propose a new scheme for PSO-based method, which involves multiple PSO swarms, called the Hierarchical Multi-swarm PSO (HMSPSO). This HMSPSO enhances search ability for the optimal solution. It is different from another multi-group PSO-based method [10]. Several researches in literature have been proposed to improve the easily-trapped problem at local minimum by the standard PSO and its variants. The HMSPSO is with a multi-level hierarchical architecture to balance the independent search by each swarm and the cooperative search among the swarms. The proposed HMSPSO is described by the following equations.

$$\begin{aligned}\mathbf{V}_i(k+1) = w \times \mathbf{V}_i(k) + c_1 \times \zeta_1 \times \left(\mathbf{pbest}_i(k) - \mathbf{L}_i(k)\right) \\ + c_2 \times \zeta_2 \times \left(\mathbf{gbest}_{1,q}(k) - \boldsymbol{L}_i(k)\right) \\ + \cdots + c_n \times \zeta_n \times \left(\mathbf{gbest}_{j,q}(k) - \boldsymbol{L}_i(k)\right)\end{aligned} \tag{12a}$$

$$\mathbf{L}_i(k+1) = \mathbf{L}_i(k) + \mathbf{V}_i(k+1) \tag{12b}$$

where $\mathbf{V}_i(k)=[v_{i,1}(k),\ v_{i,2}(k),\ldots,\ v_{i,Q}(k)]^{\mathrm{T}}$ is the velocity of the i-th particle in k-th iteration, $\mathbf{L}_i(k)=[\ l_{i,1}(k),\ l_{i,2}(k),\ldots,\ l_{i,Q}(k)]^{\mathrm{T}}$ is the location of the i-th particle, w is the inertia weight,$\{c_p,\ p=1,2,\ldots,n\}$ are the acceleration factors, $\mathbf{gbest}_{j,q}$ is the j-th level of q-th PSO swarm, and $\{\ \zeta_p, p = 1,2,\ldots,n\}$ are random number between 0 and 1.

The least squares estimation (LSE) problem can be specified with a linear model, given as follows.

$$y = \theta_1 f_1(u) + \theta_2 f_2(u) + \cdots + \theta_m f_m(u) + \varepsilon \tag{13}$$

where y is the target, u is the input to model, $\{f_i(u),\ i=1,2,..,m\}$ are known functions of u, $\{\theta_i\ ,\ i=1,2,\ldots,m\}$ are the_unknown parameters to be estimated, and ε is the model error. Note that the parameters $\{\theta_i\ ,\ i=1,2,\ldots,m\}$ can be viewed as the consequent parameters of the proposed CNFS. Observed samples can be collected to use as training data for the proposed CNFS. The training data (TD) is denoted as follows.

$$\mathrm{TD} = \{(u_i, y_i), i = 1,2,\ldots,N\} \tag{14}$$

where $(u_i,\ y_i)$ is the i-th data pair in the form of (*input, target*).Substituting data pairs into (13), we have a set of N linear equations in matrix notation, given below.

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{15}$$

where $\boldsymbol{\theta}=[\theta_1,\ \theta_2,\ \ldots,\theta_m]^{\mathbf{T}}$, $\mathbf{y}=[y_1,\ y_2,\ \ldots,y_N]^{\mathbf{T}}$, $\boldsymbol{\varepsilon}=[\varepsilon_1,\ \varepsilon_2,\ \ldots,\ \varepsilon_N]^{\mathbf{T}}$, and $\mathbf{A}$ is the matrix formed by $\{f_i(u_j), i = 1,2,\cdots,m \text{ and } j = 1,2,\ldots,N\}$. The optimal estimator for $\boldsymbol{\theta}$ can be obtained with the recursive least squares estimator (RLSE) method, given below.

$$\mathbf{P}_{k+1} = \mathbf{P}_k - \frac{\boldsymbol{P}_k \mathbf{b}_{k+1}(\mathbf{b}_{k+1})^{\mathrm{T}}\mathbf{P}_k}{1+(\mathbf{b}_{k+1})^{\mathrm{T}}\mathbf{P}_k\mathbf{b}_{k+1}} \tag{16a}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{P}_{k+1}\mathbf{b}_{k+1}(y_{k+1} - (\mathbf{b}_{k+1})^{\mathrm{T}}\boldsymbol{\theta}_k) \tag{16b}$$

for k=0,1,…,(N-1), where $[\mathbf{b}^{\mathrm{T}}_{\mathbf{k}},\ y_k]$ is the k-th row of $[\mathbf{A},\ \mathbf{y}]$. To start the RLSE algorithm, we set $\boldsymbol{\theta}_0$ to zero vector and $\mathbf{P}_0 = \alpha\mathbf{I}$, where α must be a large value and $\mathbf{I}$ is the identity matrix.

For parameter learning, the proposed CNFS predictor is trained by the hybrid HMSPSO-RLSE learning method, where the HMSPSO is used to update the premise parameters and the RLSE is used to adjust the consequent parameters. The training procedure for the proposed HMSPSO-RLSE method is given as follows.

Step 1. Collect training data. Some portion of the data is used for training, and the rest is for testing.

Step 2. Update the premise parameters by the HMSPSO in (12a) and (12b).

Step 3. Update the consequent parameters by the RLSE in (16a) and (16b), in which the row vector $\mathbf{b}$ and the vector $\boldsymbol{\theta}$ are arranged as follows.

$$\mathbf{b}_{k+1} = [\mathbf{bb}^1(k+1) \quad \mathbf{bb}^2(k+1) \quad \cdots \quad \mathbf{bb}^K(k+1)] \tag{17}$$

$$\mathbf{bb}^i(k+1) = [\lambda^i \quad h_1(k+1)\lambda^i \quad \cdots \quad h_M(k+1)\lambda^i] \tag{18}$$

$$\boldsymbol{\theta}_k = [\boldsymbol{\tau}_k^1 \quad \boldsymbol{\tau}_k^2 \ \cdots \ \boldsymbol{\tau}_k^K] \tag{19}$$

$$\boldsymbol{\tau}_k^i = [a_0^i(k) \quad a_1^i(k) \ \cdots \ a_M^i(k)] \tag{20}$$

Step 4. Calculate the CNFS output in (10).

Step 5. Calculate the cost in MSE defined below. Note that because the time series forecasting problem is in real-valued domain, only the real part of the CNFS output is involved in MSE.

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(e(t))^2 = \frac{1}{N}\sum_{t=1}^{N}(y(t) - \text{Re}(\xi(t)))^2 \tag{21}$$

Step 6. Compare the costs of all particles. Update **pbest** and **gbest** in the multiple swarms. If stopping criteria satisfied, **gbest** is the optimal premise parameters for the CNFS and stop. Otherwise, go back to Step 2 and continue the procedure.

4 Experimentation for the Proposed Approach

Example 1- Star Brightness Time Series

The star brightness time series is used to test the proposed approach for forecasting. The dataset records the daily brightness of a variable star on 600 successive midnights [19], denoted as $\{y(t), t=1,2,\ldots,600\}$. The range of dataset is normalized to the unit interval [0, 1]. We use the first 300 samples for training and the remaining 300 samples for testing. The input vector is arranged as $\mathbf{H}(t)=[y(t\text{-}3), y(t\text{-}2), y(t\text{-}1)]^{\mathrm{T}}$ and the target is given as $y(t)$ for the predicting model. Each input linguistic variable has two complex fuzzy sets, and the grid partition is designed in the input space formed by the three variables. Thus, eight T-S fuzzy rules in (3) are designed for the CNFS predictor and the NFS predictor, respectively. The cost function is designed with MSE in (21). For parameter learning, the HMSPSO algorithm in (12a) and (12b) is used to update the antecedent parameters and the consequent parameters is adjusted by the RLSE. The settings for the HMSPSO-RLSE hybrid learning method are given in Table 1. The proposed CNFS is compared to its NFS counterpart. The Gaussian-type complex fuzzy sets in (2a) to (2c) are designed for the CNFS and the traditional Gaussian fuzzy sets in (2b) are used for the NFS. Moreover, the proposed approach is compared to other approaches in [11]. The experiments with 20 trails for each are conducted. The performance comparison in average and standard deviation is shown in Table 2. For one of the 20 trials, the response by the proposed CNFS and its prediction error are shown in Figs. (2a) and (2b).

Table 1. Settings for the HMSPSO-RLSE hybrid learning method (Example 1)

HMSPSO		RLSE	
Number of premise parameters (Dimensions of particle)	18	Number of consequent parameters	16
Swarm size	300	$\boldsymbol{\theta}_0$	Zeros(1, 9)
Initial particle position	Random in $[0, 1]^{18}$	$\mathbf{P}_0$	$\alpha\mathbf{I}$
Initial particle velocity	Random in $[0, 1]^{18}$	α	10^8
acceleration parameters $\{ c_1, c_2, c_3\}$	$\{ 2, 2, 2\}$	$\mathbf{I}$	16×16 identity matrix
Swarm number of PSO	3		

Table 2. Performance Comparison (Example 1)

Method	MSE (std)	
	Training phase	Testing phase
TSK-NFIS [11]	3.14×10^{-4} (5.90×10^{-2})	3.31×10^{-4} (6.09×10^{-2})
AR [11]	3.14×10^{-4} (5.81×10^{-2})	3.22×10^{-4} (6.01×10^{-2})
NAR [11]	3.20×10^{-4} (5.96×10^{-2})	3.12×10^{-4} (5.92×10^{-2})
Neural Net [11]	3.01×10^{-4} (5.78×10^{-2})	3.11×10^{-4} (5.91×10^{-2})
PSO-RLSE for NFS	1.99×10^{-4} (4.45×10^{-6})	3.24×10^{-4} (3.27×10^{-5})
PSO-RLSE for CNFS	1.98×10^{-4} (1.03×10^{-5})	2.80×10^{-4} (1.95×10^{-5})
HMSPSO-RLSE for CNFS	1.98×10^{-4} (9.91×10^{-6})	2.72×10^{-4} (1.79×10^{-5})
Note that the results above are based on 20 trails for average and standard deviation.		

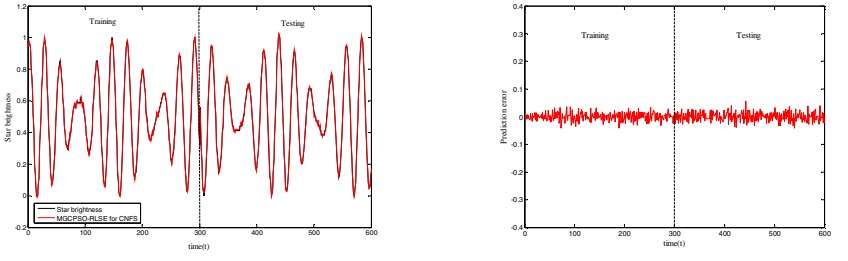


Fig. 2. (a) Prediction by the proposed approach for star brightness. (b) Prediction error.

Example 2- Oil Price Time Series

The oil price time series records the average annual price of oil, which is a small data set with 128 samples [20], denoted as $\{y(t), t=1,2,\ldots,128\}$. The range of data is normalized to the interval [0, 1]. The first 64 samples are used for training the proposed CNFS and the remaining 64 samples for testing. The input vector is arranged as $\mathbf{H}(t)=[\ y(t\text{-}2),\ y(t\text{-}1)]^{\mathrm{T}}$ and the target is given as $y(t)$ for the predicting model. Each input has two Gaussian type complex fuzzy sets in (2a) to (2c). There are four T-S fuzzy rules in (3) for the CNFS and its NFS counterpart, respectively. The CNFS has

12 premise parameters and 12 consequent parameters. The premise parameters are updated by the HMSPSO algorithm and the consequent parameters are tuned by the RLSE, as stated previously. The cost function is designed with the concept of MSE in (21). Each experiment is performed with 20 trails, and the performance comparison in average and standard deviation is shown in Table 3.

Table 3. Performance Comparison (Example 2)

Method	MSE (std)	
	Training phase	Testing phase
TSK-NFIS [11]	4.31×10^{-3} (3.42×10^{-1})	3.31×10^{-2} (6.29×10^{-1})
AR [11]	5.45×10^{-3} (3.84×10^{-1})	3.22×10^{-2} (6.38×10^{-1})
NAR [11]	4.99×10^{-3} (3.68×10^{-1})	3.12×10^{-2} (7.39×10^{-1})
Neural Net [11]	4.69×10^{-3} (3.56×10^{-1})	3.11×10^{-2} (6.50×10^{-1})
PSO-RLSE for NFS	1.98×10^{-3} (1.52×10^{-4})	2.59×10^{-2} (3.27×10^{-2})
PSO-RLSE for CNFS	2.03×10^{-3} (4.29×10^{-4})	1.63×10^{-2} (5.44×10^{-3})
HMSPSO-RLSE for CNFS	2.21×10^{-3} (2.20×10^{-4})	1.34×10^{-2} (1.34×10^{-3})
Note that the results above are based on 20 trails for average and standard deviation.		

5 Discussion and Conclusion

The proposed complex neuro-fuzzy system (CNFS) with complex fuzzy sets has been presented for the problem of time series forecasting. The CNFS is trained by the newly devised HMSPSO-RLSE hybrid learning method for the purpose of accurate forecasting. Two examples has been demonstrated for time series forecasting, and the proposed approach has shown very excellent prediction performance. This confirms our thought that the property of complex fuzzy sets (CFSs) designed into the proposed system can augment the functional mapping ability in forecasting accuracy. The notion of CFS is in contrast with that of standard type-1 fuzzy set in membership depiction. It is clearly contrasted that the membership for elements belong to a CFS is characterized within the complex-valued unit disc of the complex plane while the membership of a type-1 fuzzy set is within the real-valued unit interval between 0 and 1. Based on this contrasted property, the CNFS computing paradigm designed with complex fuzzy sets can expand the mapping flexibility for predication capability in terms of forecasting accuracy. For parameter learning of the proposed CNFS, with the divide-and-conquer concept we separate the system parameters into two smaller subsets for If-part and Then-part parameters, and then the HMSPSO-RLSE is used to train the proposed system for fast learning purpose. The idea is that the smaller the search space the easier and faster the solution can be found. This has been justified with the experiments in the two examples. In this newly devised hybrid learning method, we have implemented the multi-swam-based PSO with the RLSE algorithm, showing very good performance in terms of fast learning convergence and forecasting accuracy. For performance comparison, the proposed approach has been compared to the NFS counterpart and other approaches, the results, shown in Tables 2 and 3, show that the proposed CNFS is superior to its NFS counterpart (trained by the hybrid PSO-RLSE method). Moreover, for the forecasting accuracy, the CNFS predictor outperforms the compared approaches. Through this study, the excellence of the proposed CNFS computing approach to time series forecasting has been exposed.

Acknowledgment

This research work is supported by the National Science Council, Taiwan, ROC, under the Grant contract no. NSC99-2221-E-008-088.

References

1. Jang, S.R.: ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23, 665–685 (1993)
2. Herrera, L.J., Pomares, H., Rojas, I., Guillén, A., González, J., Awad, M., Herrera, A.: Multigrid-based fuzzy systems for time series prediction: CATS competition. Neurocomputing 70, 2410–2425 (2007)
3. Boyacioglu, M.A., Avci, D.: An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange. Expert Systems with Applications 37, 7908–7912 (2010)
4. Deng, X., Wang, X.: Incremental learning of dynamic fuzzy neural networks for accurate system modeling. Fuzzy Sets and Systems 160, 972–987 (2009)
5. Mousavi, S.J., Ponnambalam, K., Karray, F.: Inferring operating rules for reservoir operations using fuzzy regression and ANFIS. Fuzzy Sets and Systems 158, 1064–1082 (2007)
6. Ramot, D., Milo, R., Friedman, M., Kandel, A.: Complex fuzzy sets. IEEE Transactions on Fuzzy Systems 10, 171–186 (2002)
7. Dick, S.: Toward complex fuzzy logic. IEEE Transactions on Fuzzy Systems 13, 405–414 (2005)
8. Moses, D., Degani, O., Teodorescu, H.N., Friedman, M., Kandel, A.: Linguistic coordinate transformations for complex fuzzy sets. Fuzzy Systems Conference Proceedings 3, 1340–1345 (1999)
9. Ramot, D., Milo, R., Friedman, M., Kandel, A.: Complex fuzzy logic. IEEE Transactions on Fuzzy Systems 11, 450–461 (2003)
10. Niu, B., Zhu, Y., He, X., Wu, H.: MCPSO: A multi-swarm cooperative particle swarm optimizer. Applied Mathematics and Computation 185, 1050–1062 (2007)
11. Graves, D., Pedrycz, W.: Fuzzy prediction architecture using recurrent neural networks. Neurocomputing 72, 1668–1678 (2009)
12. Castro, J.L.: Fuzzy logic controllers are universal approximators. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 25, 629–635 (1995)
13. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366 (1989)
14. Buckley, J.J.: Fuzzy complex numbers. Fuzzy Sets and Systems 33, 333–345 (1989)
15. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science (1995)
16. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks Proceedings (1995)
17. Yuhui, S., Eberhart, R.C.: Fuzzy adaptive particle swarm optimization. In: Proceedings of the 2001 Congress on Evolutionary Computation (2001)
18. Mansour, M.M., Mekhamer, S.F., El-Kharbawe, N.E.S.: A Modified Particle Swarm Optimizer for the Coordination of Directional Overcurrent Relays. IEEE Transactions on Power Delivery 22, 1400–1410 (2007)
19. Time Series Data Library, Physics, Daily brightness of a variable star, `http://www-personal.buseco.monash.edu.au/hyndman/TSDL/S`
20. Time Series Data Library, Micro-Economics, Oil prices in constant 1997 dollars, `http://www-personal.buseco.monash.edu.au/hyndman/TSDL/S`

Ensemble Dual Algorithm Using RBF Recursive Learning for Partial Linear Network

Afif bin Md. Akib*, Nordin bin Saad, and Vijanth Sagayan Asirvadam

Department of Electrical & Electronic Engineering, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia

Abstract. There are many ways for gas (or high-pressure hazardous liquid) be transferred from one place to another. However, pipelines are considered as the fastest and the cheapest means to convey such flammable substances, for example natural gas, methane, ethane, benzene, propane and etc. Unavoidably, the pipelines may be affected by interference from third parties, for example human error while under its operation. Consequently, any accidental releases of gas that may occur due to the failure of the pipeline implies the risk that must be controlled. Therefore, it is necessary to evaluate the safety of the pipeline with quantitative risk assessment. Relative mass released of the leakage is introduced as the input for the simulation model and the data from the simulation model is taken at real time (on-line) to feed into the recursive algorithms for updating the linear weight. Radial basis function (RBF) is used to define the non-linear weight of the partial linear network. A new learning algorithm called the ensemble dual algorithm for estimating the mass-flow rate of the flow after leakage is proposed. Simulations with pressure liquid storage tanks problems have tested this learning approach.

Keywords: Accidental gas released mass-flow rate, recursive algorithm and radial basis function, ensemble dual algorithm.

1 Introduction

Artificial neural networks or neural networks for short are defined as a system consisting of a set of processing elements (neurons) that are connected by connections known as synapses or weights to form fully connected networks. Neural networks are non-linear in nature as the neurons usually consist of non-linear functions. Hence they are capable of learning and identifying non-linear relationships. These attributes make neural networks the ideal non-linear modeling tool in many areas of science, including control engineering.

A radial basis function neural network (RBF Network) is one of feed forward neural networks architecture besides the commonly used multi layer preceptor (MLP), which has good generalization performance especially in the non-linear system identification [1,2]. RBF network has simpler network interpretation compared to MLP [3] thus the learning process can be revealed explicitly.

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 252–261, 2011.

The term learning method is one of the most important components in any form of neural network architecture including for the RBF network. The learning or training of neural network, also known as neural computing in many research development communities, is mostly left unnoticed by many researchers whom always prefer to stick to the slow gradient decent based back-propagation, which is prone to stuck in local minima. The learning theory researchers try to find the most effective learning method to train particular network, which includes the RBF network. Learning method actually a process which tunes the parameters inside the networks so that it could represent a black-box system a process commonly known as the system identification process in control research communities, and hence it becomes one of the most important issues to be discussed in this research paper.

2 RBF Recursive Learning for Partial Linear Network

Radial basis function (RBF) is one of the special functions as its response decreases or increases monotonically with distance from the origin (center). An RBF variant or types can be any continuous function which forms the basis function in RBF network structure. There are several basis functions or kernels of RBF which may guarantee the nonlinear mapping that are needed in the RBF network learning process [4] such as Gaussian RBF, multi quadratic RBF, inverse multi quadratic RBF, thin plate splines RBF, cubic splines RBF and linear splines RBF.

RBF network consist of three layers:

- Input layer, which is made up of source nodes that connect the network to its environment.
- Hidden layer, which applies a non-linear transformation from the input layer to the last layer in RBF network (termed as output layer).
- Output layer, which supplies the response of the network to the activation pattern feed through the input layer. Hidden layer of RBF network will be characterized as non-linear parameters which consist of the center and width of the basis function.

Mathematically, RBF network with Gaussian as the activation function can be described as follows: Given arbitrary distinct samples (x_i, y_i) where $x_i = [x_i, x_{i2}, \dots\dots, x_{in}]^T \epsilon\, \Re^n, y_i \epsilon\, \Re^1$. RBF network with $\widetilde{N}$ hidden nodes can be mathematically modeled as asserted in equation 1:

$$F_{\widetilde{N}}(x) = \sum_{j=1}^{\widetilde{N}} \beta_j G\,(c_i, \sigma_i, x) \tag{1}$$

Where

$$\beta_j G\,(c_i, \sigma_i, x) = \exp\left(-\sum_{i=1}^{n} \frac{(x_i - c_{ji})^2}{2\sigma_{ij}^2}\right) \tag{2}$$

2.1 Learning in RBF Network

Learning of a neural network can be viewed as a non-linear optimization problem in which the goal is to find a set of network parameters minimizing the cost function (error function) for a given example. In other words, learning is an optimization process that produces an output that is as closed as possible to the real output of the system by adjusting the network parameters such as center, width and weight.

RBF network with Gaussian consist of two learning processes. The first is learning of the center and the second the learning of the weight. Mathematically, learning RBF network can be described as follow. Let a vector valued function as the activation as stated at equation 2 and the weight vector is (equation 3).

$$\beta = [\, \beta_1, \beta_2 \,, \ldots \ldots \,, \beta_{\tilde{n}} \,] \tag{3}$$

The overall input-output relationship of *n* input and 1-output can be described by the following non-linear relationship in equation 4:

$$\hat{y}(t, c, \sigma, \beta, x) = \sum_{j=1}^{\tilde{N}} \beta_j G\,(c_i, \sigma_i, x) = f(G, \beta, t\,) \tag{4}$$

where $\hat{y}\,(x)$ is the network output or the prediction output, c is the center of Gaussian kernel and σ is the width, β is the linear in parameter weights connection hidden and output nodes and x is the input to the network. Training or learning module of RBF network involves supplying the network with the input, determining and updating the parameter such as center, width and weight. The aim is that the networks output $\hat{y}(t, c, \sigma, \beta, x)$ approximate the desired output *y(t,x)* or the difference between the network output and the desired output are kept minimum which leads to optimization problem [5], which can be described mathematically as equation 5:

$$e = (\hat{y} - y)^2 \tag{5}$$

K-Means Clustering Method

In RBF neural network, center placement plays an important role in the learning process. Efficient center placement prior to learning process will give very significant effect on the accuracy of the RBF network model. Thus, modifying the initialization methods become the core issues.

K-means clustering is one of the most popular unsupervised learning methods used for grouping data points [6]. K-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. The K-means has been widely accepted because of its simplicity

but ability to produce a good result. Given an initial set of k-means $m_1^{(1)}, \ldots m_k^{(1)}$, which may be specified randomly or by heuristic, the algorithm proceeds in two steps:

Assignment Step: Assigning each observation to the cluster with the closest mean (equation 6).

$$S_i^{(t)} = \left\{x_j : \left\|x_j - m_i^t\right\| \leq \left\|x_j - m_{i*}^t\right\| for\ all\ i* = 1, \ldots, k\right\} \tag{6}$$

Update Step: Calculating the new mean to be the centroid of the observation in the cluster (equation 7).

$$m_i^t = \frac{1}{S_i^{(t)}} \sum_{x_j \epsilon\, S_i^{(t)}} x_j \tag{7}$$

The algorithm is deemed to have converged when the assignments no longer change. As it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum and the result is much dependent on the initial clusters.

2.2 RBF for Predicting NARX Partial-Linear Network

In partial-linear network system there are two types of weighting function; linear and non-linear weighting functions. The system is called partial-linear system when the updating process is only been carried out on the linear side of the system whereas on non-linear part, the values are predefined earlier using clustering techniques and nearest neighbor rules.

There are two types of weight parameters in the Gaussian kernel function: linear weight and non-linear weight parameters. In partial linear network, only linear weights have been updating recursively. As can be seen from equation 8, that the linear weight parameters are represented by h_i, while c and w are the non-linear weight parameters representing the center and width of the system and in partial-linear system this values are predefined earlier before the learning process is actually started.

$$\varphi(x) = \sum h_i \frac{\|x_i - c_i\|^2}{w_i^2} + h_0 \tag{8}$$

K-Means Clustering Method

The non-linear weights in partial linear network are width, w and center, c of the system. The center, c is predefined by using k-means clustering method where first of all the center of the radial basis function is initialized to be scaled according to the input size. Then the total Euclidian distance is initialized and old Euclidian distance is normalized to one.

In the beginning k-means clustering process of the center is conducted to group the data (inputs) into different form of groups and the centers are calculated. In this case, the termination criteria used is when the difference between two points (Euclidian distance) is less than $1.0e^{-14}$ and cluster of center is formed. In the case of defining the width, *w* of the system, nearest neighbor rules is engaged. The centers, *c* are used as one of the input data to determine the values of the width of the system. It begins by finding the nearest actual data points which are the closest to the centers. The next process is to find the Euclidian distance between the actual system outputs at the actual data centers and the result is sorted based on its values.

The next following step is to calculate the unit angle between the outputs and the nearest two centers and the number of nearest neighbors to use for each center. Finally, the width is calculated by taking the mean value of two neighbors.

3 Recursive Learning Algorithms for Updating Linear Weights

The linear weight in this context is the heights of the RBF network. In the beginning the height is initialized as random based on the number of the hidden neuron plus bias and number of output. Based on the number of input data into the system, the covariance matrix (P_{mat}) is calculated and updated each time. The height of RBF network is also updated according to the value of the covariance matrix and this can be shown in equation 9.

$$H = H + Lr * P_{mat} * Input * error \tag{9}$$

where *Lr* is the learning rate which is set to be 0.95 and *error* is the difference between the system output and estimate output of the system. There are three methods used for updating the covariance matrix in the case of updating the linear weight of partial linear network. The three methods are:

- Recursive least square (RLS) method.
- Recursive levenberg marquadt (RLM).
- Ensemble dual algorithm method.

RLS algorithm can be expressed as:

$$\theta_t = \theta_{t-1} + P_t \phi_t e_t \tag{10}$$

$$e_t = y_t + \phi_t^T \theta_{t-1} \tag{11}$$

$$P_t = P_{t-1} - \frac{P_{t-1}\phi_t \phi_t^T P_{t-1}}{1+\phi_t^T P_{t-1}\phi_t} \tag{12}$$

RLM algorithm can be expressed as:

$$S = \gamma \wedge_t + \ \Omega^t P_{t-1} \Omega \tag{13}$$

$$P_t = \frac{1}{\gamma}[P_{t-1} - P_{t-1}\Omega S^{-1}\Omega^T P_{t-1}] \tag{14}$$

$$\theta_t = \theta_{t-1} + P_t \phi_t e_t \tag{15}$$

3.1 Ensemble Dual Recursive Learning

Ensemble dual recursive learning algorithm is a new proposed algorithm which is a combination of any two recursive algorithms. Ensemble dual algorithm as proposed in this paper, work based on two methods: based on the weighting function and another is based on binary selection method.

This algorithm is used to predict the mass flow rate of a flow after any accidental dispersion occurs (prediction of Q_{out}) in particular and prediction of any system in general. In ensemble dual algorithm, there are two sets of covariance matrices, each one is a respective to each algorithm employed and one main covariance matrix, make it all together three covariance matrices involved. Not only that, each covariance matrices will correspond to produce its own prediction error matrices and separate linear weight matrices, θ_t.

3.1.1 Weighting Function Method

In order to show how the ensemble dual algorithm works, an example of the ensemble dual recursive training using RLS & RLM algorithms is as shown. There are two covariance matrices involves respectively, the RLS and RIV algorithms:

$$P_{RLSt} = \frac{1}{\gamma}\left(P_{RLSt-1} - \frac{P_{RLSt-1}\phi_t \phi_t^T P_{RLSt-1}}{\gamma + \phi_t^T P_{RLSt-1}\phi_t}\right) \tag{16}$$

$$P_{RLMt} = \frac{1}{\alpha_t}[P_{RLMt-1} - P_{RLMt-1}\Omega(w_t)S^{-1}\Omega^T(w_t)P_{RLMt-1}] \tag{17}$$

These covariance matrices are then used to update the value of weights estimation (θ_t) separately according to its own algorithm and at the same time calculating the prediction error as follows:

RLS:

$$\theta_{RLSt} = \theta_{RLSt-1} + P_{RLSt}\phi_t e_{RLSt} \tag{18}$$

$$e_{RLSt} = y_t + \phi_t^T \theta_{RLSt-1} \tag{19}$$

RLM:

$$\theta_{RLMt} = \theta_{RLMt-1} + P_{RLMt} z_t e_{RLMt} \tag{20}$$

$$e_{RLMt} = y_t + \phi_t^T \theta_{RLMt-1} \tag{21}$$

From these weighting matrices and estimation error values, the main weighing matrices, θ_{Mt} is calculated. The main θ_{Mt} is calculated based on the mean square errors (MSE) which are calculated earlier. The formulation of the main θ_{Mt} is given as follows:

$$\theta_{Mt} = \theta_{Mt-1} + \left[\frac{MSE_{RLS}}{MSE_{RLS} + MSE_{RLM}}\right] P_{RLS}\phi_t e_{RLS} + \left[\frac{MSE_{RLM}}{MSE_{RLS} + MSE_{RLM}}\right] P_{RLM}\phi_t e_{RLM} \quad (22)$$

One formulated the main θ_{Mt} can be used to calculate the prediction value of the mass flow rate of the flow after any accidental dispersion occurs (prediction value of Q_{out}). Mean square error is once again calculated with respect to the main θ_{Mt}.

3.1.2 Binary Selection Method

Similar to the weighting function method, in binary selection, there are also two sets of covariance matrices: which consist of two sets of weighting function and two sets of prediction error plus the main θ_{Mt}. What makes this method different is that every time prediction error RLS and RLM are calculated (equation 19 and 21) this algorithm will be compared between these two values and will take the weighting function from the smaller value to be one of the elements of the matrices main θ_{Mt}. Example of how this method works is presented in the following equations.

If $(e_{RLSt} > e_{RLMt})$

$$\theta_{Mt} = \theta_{Mt-1} + P_{RLMt} z_t e_{RLMt} \quad (23)$$

Else

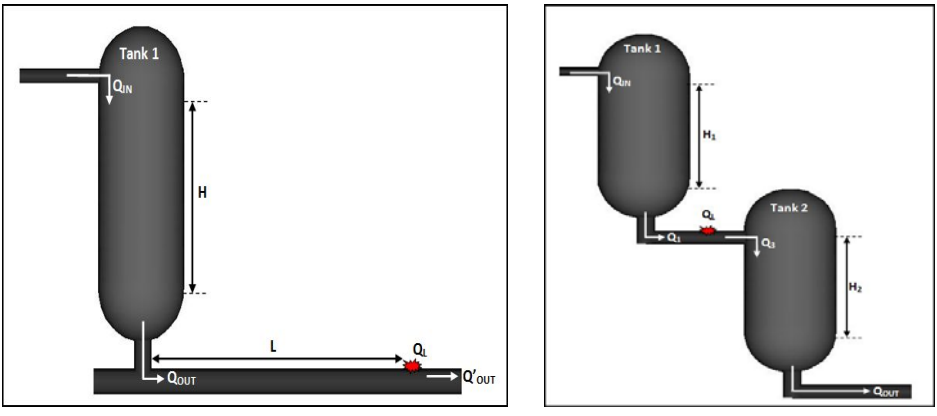
$$\theta_{Mt} = \theta_{Mt-1} + P_{RLSt}\phi_t e_{RLSt} \quad (24)$$

As stated in equations (23) and (24), the θ_{Mt} will be taken from both algorithms which is selected based on the value of the error generated between RLS and RLM algorithm at a time. Then the main θ_{Mt} is used to calculate the mass prediction error (prediction Q_{out}) and following this the mean square error can then be calculated.

4 Case Studies

In real process industries, there are many types of tanks and pipes organised in many different ways. The tanks and pipes are varied from one to another based on its functionality, sizes, organisation, flow parameters and etc. This work gives focus on the problems that deal with high pressure liquid storage tanks connected with series of pipes and tanks. There are two case studies designed and being considered and for each case study there is one leakage at a particular point. The two case studies are:

a. Single Tank – single input and single output (Figure 1 (a)).
b. Two Tanks – single input and single output (coupled tanks) (Figure 1 (b)).



(a) (b)

Fig. 1. (a) A single tank system (b) A Coupled Tanks system

5 Results and Discussions

There are four different sets of recursive algorithms being tested for the two case studies, namely single tank and couple tank. These algorithms are the RLS, RLM ($\rho = 0.0001$) and RLS-RLM. For each setting, twenty-five simulations were carried out and the average or means of the results are tabulated in Tables 1 and 2.

Case Study1: Single Tank

In the case of single tank – single input and single output as depicted in Figure 1(a), the mass flow after the leakage is considered and to be estimated using the different algorithms. Comparing the two methods in ensemble dual algorithm, binary selection method always gives better prediction as compared to weighting function method.

On overall for the single tank single input and single output, binary selection RLS-RLM ($\rho = 0.0001$) gives the best estimation with very small mean prediction error. In general, RLS and RLM always give good estimation and can be improved by using the ensemble dual method.

Case Study 2: Coupled Tank

The second test case is on a coupled tank, with single input and single output arrangement as depicted in Figure 1(b). In this case, on-line system identification is used to estimate the mass flow-rate of the flow from the second tank, where the leakage is made to occur before the flow goes from tank 1 to tank 2.

The same algorithm as used in the case of the single tank is also being used here. The aim is to evaluate whether the same algorithms can be used to estimate the mass flow-rate for the case of the coupled tank. There is a consistency in the results for

both cases, the single tank and the coupled tanks, in which the ensemble RLS-RLM gives the best overall result. In other words, the methods considered in the studies are shown to be effective to estimate the mass flow-rate after leakage with small mean prediction error. Nevertheless, at the beginning of the estimation process for all the methods, the systems are not being able to estimate the flow very well. There are always overestimate of the flow, which lead to large mean prediction errors. Taken the last 400 values of each method, the results show that all methods can predict the flow very well and the mean prediction error is smaller as compared to the overall mean prediction error.

The results for both case studies can be seen in the tables 1 and 2 below:

Table 1. Numerical Result for Partial Linear Single Tanks with Number of Cycle (T_c) = 500 and Hidden Neurons (N_h) varies from 5 to 15

Recursive Algorithm Structure			Performance Measures				
Number of Cycle Time	Number of Hidden Neurons	Algorithms	Mean Square Error (MSE)	Standard Deviation MSE	MSE (last 400)	Max MSE (Last 400)	Standard Deviation MSE (last 400)
500	5	RLS	1.4355	10.0295	0.2192	8.5946	0.9299
		RLM (ρ=0.0001)	0.0581	0.4832	0.0026	0.0442	0.0054
		RLS – RLM (weighting)	2.2786	15.8714	0.0971	1.1018	0.1807
		RLS – RLM (binary)	0.6982	5.1209	0.1071	0.5903	0.0052
500	10	RLS	6.9873e+07	9.400e+08	1.424e+08	3.569e+09	5.182e+08
		RLM (ρ=0.0001)	0.0116	0.0433	0.0109	0.3630	0.0331
		RLS – RLM (weighting)	7.873e+02	4.813e+4	8.591e+03	3.8415e+05	6.9144e+06
		RLS – RLM (binary)	0.0696	0.4994	0.0023	0.0530	0.0051
500	15	RLS	0.0179	0.1166	0.0137	0.7530	0.0646
		RLM (ρ=0.0001)	0.0169	0.1165	0.0036	0.1273	0.0123
		RLS – RLM (weighting)	**0.0098**	**0.0757**	**3.3050e-04**	**0.0209**	**0.0015**
		RLS – RLM (binary)	0.5303	3.9913	0.0432	0.7120	0.1069

Table 2. Numerical Result for Partial Linear Coupled Tank with Number of Cycle (T_c) = 500 and Hidden Neurons (N_h) varies from 5 to 15

Recursive Algorithm Structure			Performance Measures				
Number of Hidden Neurons	Number of Hidden Neurons	Algorithms	Mean Square Error (MSE)	Standard Deviation MSE	MSE (last 400)	Max MSE (Last 400)	Standard Deviation MSE (last 400)
500	5	RLS	0.0523	0.2720	0.0312	2.9749	0.1790
		RLM (ρ = 0.0001)	8.4969	36.4702	0.0201	0.3032	0.0447
		RLS – RLM (weighting)	0.3217	1.4269	0.5460	17.0202	1.9511
		RLS – RLM (binary)	0.9431	0.6712	0.0151	1.5701	0.08549
500	10	RLS	0.1174	0.6115	0.0048	0.2157	0.0189
		RLM (ρ = 0.0001)	0.0304	0.2041	0.0057	0.0745	0.0106
		RLS – RLM (weighting)	0.0442	0.2430	0.0301	0.4250	0.0690
		RLS – RLM (binary)	0.0258	0.0451	0.0035	0.0612	0.0073
500	15	RLS	0.0056	0.0432	1.636e-04	0.0027	3.5834e-04
		RLM (ρ = 0.0001)	0.0051	0.0344	4.0107e-04	0.0084	9.0668e-004
		RLS – RLM (weighting)	0.0101	0.0477	3.6800e-04	0.0049	5.9120e-004
		RLS – RLM (binary)	**0.0032**	**0.0068**	**7.912e-05**	**0.0017**	**5.6783e-04**

6 Conclusion

This work investigates the performances of RBF for defining the non-linear weights and recursive algorithms and their capabilities to estimate the mass flow-rate of various types of tanks with the presence of leakage. In general, RLS and RLM algorithms show promising and consistent results.

By combining two recursive algorithms namely RLS and RLM, a new algorithm called the ensemble dual algorithm is proposed. To illustrate its applicability, two methods of updating its covariance matrix; the first is the weighting function method and the second is the binary selection method. In the two cases tested, ensemble dual algorithm with binary selection method has proved to be the best overall method for estimating the mass flow-rate of the flow. The main reason behind the proposal of this algorithm is to provide an alternative algorithm that inherits only the best prediction of the two algorithms combined.

References

1. Surandrajan, N., Saratchandran, P., Wei, L.Y.: Radial Basis Function Neural Networks with Sequential Learning, MRAN and Its Application. In: Progress in Neural Processing, vol. 11, World Scientific Publishing Co. Pte. Ltd., Singapore (1999)
2. Huang, G.B., Siew, C.K.: Extreme Learning Machine with Randomly Assigned RBF Kernels. International Journal of Information Technology 11(1), 16–24 (2004)
3. Chi, F.F., Billings, S.A., Luo, W.: On-Line Supervised Adaptive Training Using Radial Basis Function Networks. In: Neural Network, vol. 19, pp. 1597–1617. Elsevier Science, Amsterdam (1996)
4. Gupta, M., Jin, L., Humma, N.: Static and Dynamic Neural Networks. Wiley Interscience, Hoboken (2003)
5. Ngia, S.H., Sjoeberg, J.: Efficient training of neural nets for nonlinear adaptive filtering using a recursive Levenberg-Marquadt algorithm. IEEE Transactions on Signal Processing 48, 1915–1927 (2000)
6. Isaksson, A.J.: Identification of ARX-Models Subject to Missing Data. IEEE Transaction on Automatic Control 35(5), 813–819 (1993)
7. Asirvadam, V.S., McLoone, S.F., Irwin, G.W.: Parallel and separable recursive Levenberg-Marquardt training algorithm. In: Proceedings of the 2002 Neural Networks, pp. 129–138 (2002)
8. Yuhua, D., Huilin, G., Jing'en, Z., Yaorong, F.: Evaluation of gas release rate through holes in pipelines. Journal of Loss Prevention in the Process Industries 15(6), 423–428 (2002)
9. Du, K.L., Swammy, M.N.S.: Neural Networks in a Soft computing Framework, Spinger, London (2006)
10. Jun, Y., Meng Er., J.: An Enhanced On-line Sequential Extreme Learning Machine. In: IEEE Control and Decision Conference, pp. 2902–2907. IEEE, Los Alamitos (2008)
11. Zan, L., Latini, G., Pollini, G., Baldelli, P.: Landslides early warning monitoring system. In: Geosciences and Remote Sensing Symposium, vol. 1, pp. 188–190 (2002)

A Novel Hybrid Forecast Model with Weighted Forecast Combination with Application to Daily Rainfall Forecast of Fukuoka City

Sirajum Monira Sumi*, Md. Faisal Zaman, and Hideo Hirose

Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
sumi@ume98.ces.kyutech.ac.jp,
zaman@ume98.ces.kyutech.ac.jp,
hirose@ces.kyutech.ac.jp

Abstract. In this paper, we propose a novel hybrid multi-model approach for rainfall forecasting. In this multi-model system we have incorporated an efficient input selection technique, a set of distinct predictive models with carefully selected parameter settings, a variable selection method to rank (weight) the models before combining their outputs and a simple weighted average to combine the forecasts of all the models. The input selection technique is based on auto correlation and partial autocorrelation function, the predictive models are stepwise linear regression, partial least square regression, multivariate adaptive regression spline, radial basis kernel gaussian process and multi layer perceptron with quasi Newton optimization. The model ranking technique is based multi response sparse regression, which rank the variables (here models) according to their predictive performance (here forecasting). We have utilized this rank to use it as the *wegiht* in the weighted average of the forecast combination of the models. We have applied this novel multi model approach in forecasting daily rainfall of rainy season of Fukuoka city of Japan. We have used several performance metrics to quantify the predictive quality of the hybrid model. The results suggest that the novel hybrid multi-model approach can make efficient and persistent short term rainfall forecast.

Keywords: input selection, variable ranking, weighted forecast combination, daily rainfall, short term forecast.

1 Introduction

Accurate information concerning the amount of rainfall is essential for the use and management of water resources. More specifically in the cities, rainfall has a strong effect on traffic control, the operation of sewer systems, and other human activities. It should also be noted that, rainfall is one of the most complex and difficult component of the hydrology cycle to decipher and also to model

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 262–271, 2011.

due to the tremendous range of variation over a wide range of scales both in space and time. The intricacy of the atmospheric processes that generate rainfall makes quantitative forecasting of rainfall an extremely difficult task. Thus, to construct a predictive system to produce accurate rainfall forecasting is one of the greatest challenges for the researchers of diverse fields such as weather data mining [19], environmental machine learning [5], operational hydrology [8], statistical forecasting [10], despite many advances in weather forecasting in recent decades. The parameters that are required to predict rainfall are enormously complex and subtle even for a short time period.

Recently, the concept of coupling different models has attracted more attention in hydrologic forecasting. They can be broadly categorized into ensemble models and modular (or hybrid) models. The basic idea behind ensemble models is to build several different or similar models for the same process and to integrate them together [18]; [1]; [6]). For example, Xiong [18] used a Takagi-Sugeno-Kang fuzzy technique to couple several conceptual rainfall-runoff models. Coulibaly et al. (2005) employed an improved weighted-average method to coalesce forecasted daily reservoir inflows from K-NN, conceptual model and ANN. Kim et al. [6] investigated five ensemble methods for improving stream flow prediction.

Physical processes in rainfall are generally composed of a number of subprocesses. Their accurate modeling by building of a single global model is sometimes not possible [14]. Modular models were therefore proposed where sub-processes were first of all identified (or without identifying) and then separate models (also called local or expert model) were established for each of them [14]. Different modular models were proposed depending on the soft or hard splitting of training data. Soft splitting means the dataset can be overlapped and the overall forecasting output is the weighted average of each local model [12]; [17]. On the contrary, there is no overlap of data in the hard splitting and the final forecasting output is explicitly from only one of local models [17]. Our approach in this paper is related with soft splitting of the data.

In this paper we have developed a hybrid (or multi-model) forecast model with an appropriate input selection technique coupled with appropriate data-preprocessing technique and model selection (for weighted forecast combination) to improve the accuracy of rainfall forecasting while using observed rainfall records in both space and time. In order to overcome the problem encountered in training a linear model we have used five distinctive alternative models (both linear and non-linear). Also we have designed each of the models with different design (i.e., with different values for the parameters). Using this hybrid model, we have forecasted rainfall of Fukuoka city from 1 day ahead, using continuous daily rainfall data of rainy season (June and July) from 1975 to 2010.

The rest of the paper is follows, in Section 2 we have discussed briefly about the study area and the rainfall series used in this paper. In Section 3 we have described the hybrid forecast model including the input selection technique and the variable selection method and how the weights are extracted. This is followed by discussions about the experimental setup (Section 4) and results (Section 5). Lastly conclusive discussions of the paper is in Section 6.

2 Study Area

In this paper we have taken the daily rainfall series of rainy season from nearby weather stations of Fukuoka city to forecast the rainfall of Fukuoka city in rainy season. Each weather station is within the range of 48 km from Fukuoka city. Considering the distance the rainfall data is taken taken from 6 forecast stations (as forecast point) in Fukuoka and Saga prefecture in Japan. We plotted the rainfall series in Figure 1. Our objective is to forecast 1-day ahead rainfall of rainy season in Fukuoka city.

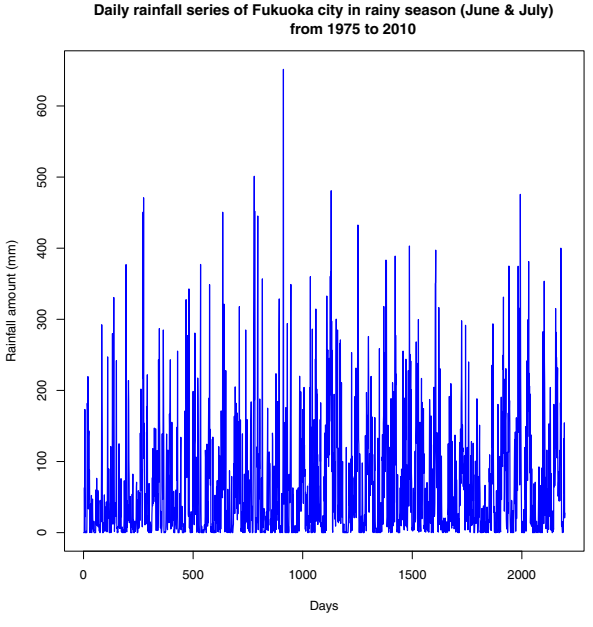


Fig. 1. Daily rainfall series of Fukuoka city in rainy season (June & July) from 1975 to 2010

3 Hybrid Forecast Model

In this section we have described the proposed hybrid forecast model. A hybrid (or multi-model) forecast model with an appropriate input selection technique coupled with appropriate data-preprocessing technique and model selection (for weighted forecast combination) is proposed. The model selection has been inserted to improve the accuracy of rainfall forecasting. Rainfall is a complex stochastic process comprising linear and non-linear phenomenon; in order to extract better learning of the process, we have used five distinctive alternative models (both linear and non-linear). Also we have designed each of the models with different design (i.e., with different values for the parameters). The architecture of the forecast model is presented in Figure 2.

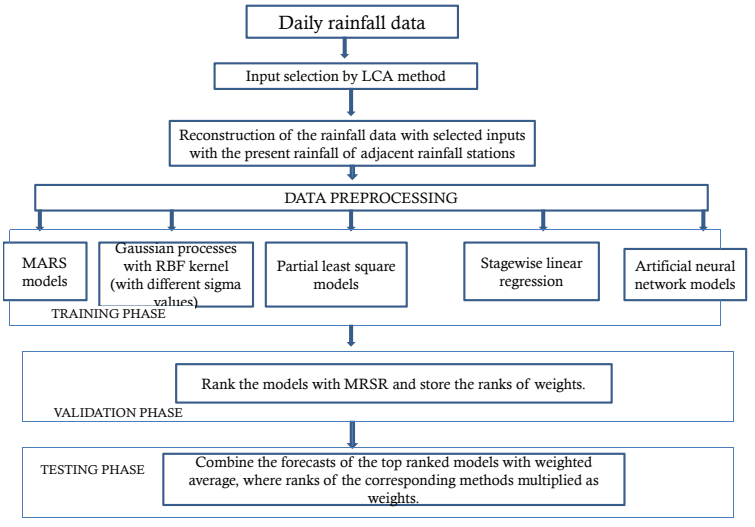


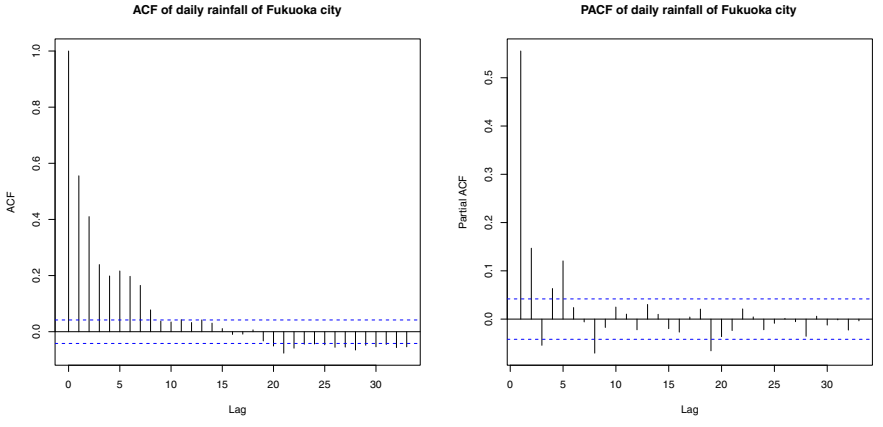
Fig. 2. Architecture of the Hybrid forecasting model

3.1 Determination of Model Inputs

We have used the Linear Correlation Analysis (LCA) [15] method for selecting the proper inputs from the rainfall series, which in other way can be stated as selecting the proper lag value which can appropriately represent the previous values. In LCA the value autocorrelation function (ACF) and partial autocorrelation function (PACF) would suggest the influencing previous patterns in the series at a given time. The ACF and PACF with 95% confidence levels is examined, and the number of previous rainfall values for which both ACF and PACF has high value should be included in the input vector. The variables that may not have a significant effect on the performance of the model can be trimmed off from the input vector, resulting in a more compact model. In our data we see that at lag = 2 (see Figure 3) the values of both ACF and PACF is highest, so we shall consider rainfall inputs with lag = 2.

3.2 Data Preprocessing

Moving Average (MA). As our model is based on soft splitting the data, so we have used MA to split the data into over lapping parts. MA smoothes data by replacing each data point with the average of the k neighboring data points, where k may be termed the length of memory window. The method is based on the idea that any large irregular component at any point in time will exert a smaller effect if we average the point with its immediate neighbors. The equally weighted MA is the most commonly used, in which each value of the data carries the same weight in the smoothing process.



(a) Auto-correlation function (ACF) (b) Partial ACF (PACF)

Fig. 3. ACF and PACF values of daily rainfall at different lags. We can see at lag = 5 the values of both ACF and PACF is highest than other lags. The dotted line is the 95% confidence interval.

Data Scaling. It is worthwhile to notice that the scaling of the training data is very crucial in the improvement of the model performance. The scaling as adopted above for model input determination, is to scale the training data to $[0, 1]$ or even more narrow interval. We have used the following formula to scale the data to interval $[0, 1]$

$$Z = \frac{y - \min(y)}{\max(y) - \min(y)}$$

3.3 Multi Models

Partial Least Square Regression (PLS). Partial least squares (PLS) regression is used to describe the relationship between multiple response variables and predictors through the latent variables. PLS regression can analyze data with strongly collinear, noisy, and numerous X-variables, and also simultaneously model several response variables, Y. It is particularly useful when we need to predict a set of dependent variables from a large set of independent variables (i.e., predictors). In technical terms, PLS regression aims at producing a model that transforms a set of correlated explanatory variables into a new set of uncorrelated variables. The parameter coefficients between the predictor variables and the criterion variable.

Gaussian Process. Gaussian process regression [11] is a nonparametric method based on modeling the observed responses of the different training data points (function values) as a multivariate normal random variable. For these function

values an a priori distribution is assumed that guarantees smoothness properties of the function. Specifically, the correlation between two function values is high if the corresponding input vectors are close (in Euclidean distance sense) and decays as they go farther from each other. The posterior distribution of a to-be-predicted function value can then be obtained using the assumed prior distribution by applying simple probability manipulations.

Multivariate Adaptive Regression Spline (MARS). The MARS model is a spline regression model that uses a specific class of basis functions as predictors in place of the original data [4]. The MARS basis function transform makes it possible to selectively blank out certain regions of a variable by making them zero, allowing MARS to focus on specific sub-regions of the data. MARS excels at finding optimal variable transformations and interactions, as well as the complex data structure that often hides in high-dimensional data.

Stepwise Linear Regression (SLR). The linear regression model herein is actually called stepwise linear regression (SLR) model because the forward stepwise regression is used to determine optimal input variables. The basic idea of SLR is to start with a function that contains the single best input variable and to subsequently add potential input variables to the function one at a time in an attempt to improve the model performance. The order of addition is determined by using the partial F-test values to select which variable should enter next. The high partial F-value is compared to a (selected or default) F-to-enter value. After a variable has been added, the function is examined to see if any variable should be deleted.

Neural Network. The multilayer perceptron network is by far the most popular ANN paradigm, which usually uses the technique of error back propagation to train the network configuration. The architecture of the ANN consists of a number of hidden layers and a number of neurons in the input layer, hidden layers and output layer. ANNs with one hidden layer are commonly used in hydrologic modeling. In this paper we have used the quasi Newton optimization to optimize the network.

3.4 Model Selection (and Ranking) Using Multi-Response Sparse Regression (MRSR)

For the removal of the useless neurons of the hidden layer, the Multiresponse Sparse Regression proposed (MRSR) by Timo Similä and Jarkko Tikka in [13] is used. It is an extension of the Least Angle Regression (LARS) algorithm [3] and hence is actually a variable ranking technique, rather than a selection one. We have used this model ranking in the validation phase to estimate the weights of the individual models. These weights are later utilized for weighted average combination [16] of the forecasts of the models in the testing phase.

4 Experiment

In the experiment we have split the data into three parts: a) training: from 1975 to 1999, b) validation: from 2000 to 2004 and c) testing: from 2005 to 2010. In training phase we have trained all the models (except SLR) with 5 different parameter settings, so this means we have total 4x5+1 = 21 models in this phase. The parameters of the models are suitably chosen with prior knowledge of the methods. In this phase we have used MA(1) i.e, moving average with single window (based on the RMSE value during training) to soft split the data, then this data is used for training. The forecast of these methods will be stored in the forecast matrix. In the validation phase the MRSR is used on this forecast matrix to rank the methods and we select only the top 10 models (with highest weight) for testing. In the testing phase we combine the forecast of the top models with the weighted average, where the weights are (ranks of the corresponding methods) estimated in the validation phase.

Table 1. Evaluation metrics used in this paper with their perfect scores

Metric	Formula	Perfect Score
Root Mean Sum of Square Error (RMSE)	$\frac{1}{N}\sum_{t=1}^{N}(F_t - O_t)^2$	Low values better
Coefficient of Efficiency (C.E)	$1 - \frac{\sum_{t=1}^{N}(F_t-O_t)^2}{\sum_{t=1}^{N}(O_t-\bar{O})^2}$	1
Persistency Index (P.I)	$1 - \frac{\sum_{t=1}^{N}(F_t-O_t)^2}{\sum_{t=1}^{N}(O_t-O_{t-l})^2}$	1
Bias	$\frac{\sum_{t=1}^{N}(F_t)}{\sum_{t=1}^{N}(O_t)}$	1
Normalized Mean Sum of Square Error (NMSE)	$\frac{1}{N\sigma^2}\sum_{t=1}^{N}(F_t - O_t)^2$	Low values better
Structured Mean Absolute Percent Error (sMAPE)	$\frac{1}{N}\sum_{t=1}^{N}\lvert\frac{(F_t-O_t)}{(F_t-O_t)}\rvert$	Low values better
Correlation Coefficient (CC)	$\frac{\sum_{t=1}^{N}(O_t-\bar{O})(F_t-\bar{F})}{\sqrt{\sum_{t=1}^{N}(O_t-\bar{O})^2 * \sum_{t=1}^{N}(F_t-\bar{F})^2}}$	1

4.1 Evaluation of Model Performances

To evaluate the hybrid model we have used several evaluation metrics (please see Table 1). In Table 1, O_t is observed rainfall and F_t is the forecasted rainfall. We have given the perfect score of each metric so that the reader can evaluate the models properly. The coefficient of efficiency (CE) [9] is a good alternative to R^2 as a "goodness-of-fit" in that it is sensitive to differences in the observed and forecasted means and variances. The Persistence Index (PI) [7] was adopted here for the purpose of checking the prediction lag effect. Other metrics are popular so we opt not to go to the details of those.

5 Result and Discussion

In this section we have presented the results of validation and testing phase. In Figure 4 we have presented hyetograph of validation and testing phase. from these plots we see that the rainfall forecasts in both phase follow the observed the rainfall. In Figure 5 we have presented the scatter plot of observed vs the

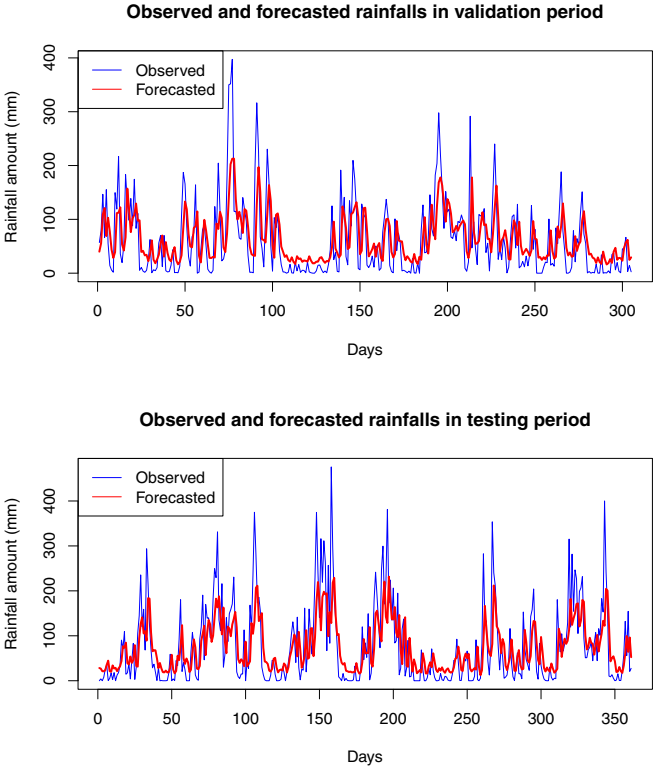


Fig. 4. The hyetograph of one step ahead forecast at validation phase (upper) and at test phase (down)

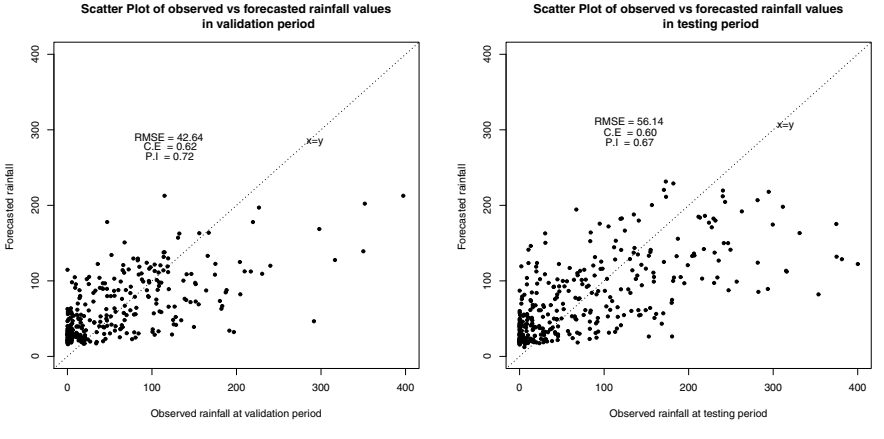


Fig. 5. Scatter plots of one step ahead forecast at validation and testing phase

forecasted rainfall values. For a perfect forecast all the points would be around the dotted line (in the middle) of both the plots. We have also inserted the RMSE, C.E and P.I values for reference.

In Table 2 we have presented the values of NMSE, Bias, sMAPE and CC. As we have mentioned in Table 1 the perfect score of the metrics we can see that all the values of each of the metrics are around satisfactory range in validation phase and testing phase.

Table 2. Forecasting accuracy of the hybrid model in terms of various evaluation metrics

Metric	Validation	Testing
Bias	0.9492	1.0520
NMSE	0.3768	0.3929
sMAPE	0.4526	0.4517
CC	0.7820	0.8235

6 Conclusion

A novel hybrid forecast model is proposed in this paper, with a data driven input selection technique, with a data driven data preprocessing method and coupled with suitable multi models ranking (and hence selected) on the basis of an appropriate variable (here model) ranking algorithm with a weighted average to combine the forecast of the top models. Then this hybrid model is applied to forecast 1-step ahead rainfall forecast of Fukuoka city in the rainy season. The performance of the hybrid model as visualized from several plots and values of performance metrics indicate that the hybrid forecasting model is capable of forecasting rainfall accurately.

References

1. Abrahart, R.J., See, L.M.: Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrology and Earth System Sciences 6(4), 655–670 (2002)
2. Coulibaly, P., Haché, M., Fortin, V., Bobée, B.: Improving daily reservoir inflow forecasts with model combination. Journal of Hydrologic Engineering 10(2), 91–99 (2005)
3. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annal. Statist. 32, 407–499 (2004)
4. Friedman, J.H.: Multivariate Adaptive Regression Splines. Annal. Statist. 19, 1–141 (1991)
5. Hong, W.C.: Rainfall forecasting by technological machine learning models. App. Math. Computat. 200, 41–57 (2008)
6. Kim, T., Heo, J.H., Jeong, C.S.: Multireservoir system optimization in the Han River basin using multi-objective genetic algorithms. Hydrological Processes 20(9), 2057–2841 (2006)

7. Kitanidis, P.K., Bras, R.L.: Real-time forecasting with a conceptual hydrologic model applications and results. Water Resources Research 16(6), 1034–1044 (1980)
8. Li, P.W., Lai, E.S.T.: Short-range quantitative precipitation forecasting in Hong Kong. J. Hydrology 288, 189–209 (2007)
9. Nash, J.E., Sutcliffe, J.V.: River flow forecasting through conceptual models part I A discussion of principles. J. Hydrology 10(3), 282–290 (1970)
10. Pucheta, J., Patino, D., Kuchen, B.: A statistically dependent approach for the monthly rainfall forecast from one point observations. In: Li, D., Chunjiang, Z. (eds.) IFIP International Federation for Information Processing, vol. 294. Computer and Computing Technologies in Agriculture II, vol. 2, pp. 787–798. Springer, Boston (2009)
11. Rasmussen, C.E., Williams, K.L.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
12. Shrestha, D.L., Solomatine, D.P.: Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 19(2), 225–235 (2006)
13. Similä, T., Tikka, T.: Multiresponse sparse regression with application to multidimensional scaling. In: Proceedings 15th International Conference on Artificial Neural Networks, Part II, pp. 97–102 (2005)
14. Solomatine, D.P., Ostfeld, A.: Data-driven modelling: some past experiences and new approaches. J. Hydroinformatics 10(1), 3–22 (2008)
15. Sudheer, K.P., Gosain, A.K., Ramasastri, K.S.: A data-driven algorithm for constructing artificial neural network rainfall runoff models. Hydrological Processes 16, 1325–1330 (2002)
16. Timmermann, A.: Forecast combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds.) Handbook of Economic Forecasting, pp. 135–196. Elsevier Pub., Amsterdam (2006)
17. Wu, C.L., Chau, K.W., Li, Y.S.: River stage prediction based on a distributed support vector regression. J. Hydrology 358, 96–111 (2008)
18. Xiong, L.H., Shamseldin, A.Y., O'Connor, K.M.: A non-linear combination of forecasts of rainfall-runoff models by the first-order TS fussy system for forecast of rainfall runoff model. J. Hydrology 245, 196–217 (2001)
19. Yang, Y., Lin, H., Guo, Z., Jiax, J.: A data mining approach for heavy rainfall forecasting based on satellite image sequence analysis. Computers & Geosciences 33, 20–30 (2005)

Estimation of Optimal Sample Size of Decision Forest with SVM Using Embedded Cross-Validation Method

Md. Faisal Zaman* and Hideo Hirose

Kyushu Institute of Technlogy
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
zaman@ume98.ces.kyutech.ac.jp
hirose@ces.kyutech.ac.jp

Abstract. In this paper the performance of the m-out-of-n decision forest of SVM without replacement with different subsampling ratio ($\frac{m}{n}$) is analyzed in terms of an *embedded cross-validation* technique. The subsampling ratio plays a pivotal role in improving the performance of the decision forest of SVM. Because the SVM in this ensemble enlarge the feature space of the underlying base decision tree classifiers and guarantees a improved performance of the ensemble overall. To ensure the better training of the SVM generally the out-of-bag sample is kept larger but there is no general rule to estimate the optimal sample size for the decision forest. In this paper we propose to use the embedded cross-validation method to select the a near optimum value of the sampling ratio. In our criterion the decision forest of SVM trained on independent samples whose size is such that the cross-validation error of that ensemble is as low as possible, will produce an improved generalization performance for the ensemble.

Keywords: Optimal sampling ratio, Decision forest with SVM, Embedded cross-validation error.

1 Introduction

Ensemble learning is one of the main research directions in recent years, due to their potential to improve the generalization performance of the predictors. It has attracted scientists from several fields including Statistics, Machine Learning, Pattern Recognition and Knowledge Discovery. Numerous theoretical and empirical studies have been published to establish the advantages of the predictor decision combination paradigm over the single (individual) predictor [12], [10]. The success of ensemble methods arises largely from the fact that they offer an appealing solution to several interesting learning problems of the past and the present, such as improving predictive performance, learning from multiple physically distributed data sources, scaling inductive algorithms to large

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 272–281, 2011.

databases and learning from concept-drifting data streams. Most popular among the ensemble creation techniques are Baggign [2], Adaboost [8], Random Forest [4], Multiboost [19] and Rotation Forest [14].

Two of the most popular classifier ensemble schemes is Bagging [2] and Adaboost [8] (the most popular) of the Boosting family [15]. In standard bagging individual classifiers are trained on independent bootstrap samples that are generated with replacement from the set of labeled training samples, where as in adaboost by *resampling* the fixed training sample size and training examples resampled according to a probability distribution are used in each iteration and in each iteration, the distribution of the training data depends on the performance of the classifier trained in the previous iteration. Unlike these two ensembles a *bagging type* hybrid ensemble method was proposed by Hothorn and Lausen, defined as *Double Bagging* [11] to add the outcomes of arbitrary classifiers to the original feature set for bagging of classification trees. Double bagging is a combination of linear discriminant analysis and classification trees. As in the bootstrap method, approximately $\frac{1}{3}$ of the observations in the original training set are not part of a single bootstrap sample in bagging [3], and Breiman termed the set constituted by these observations as an out-of-bag sample. In double bagging, the decision forest is constructed by utilizing an out-of-bag sample (OOBS) to estimate the coefficients of a linear discriminant function (LDA) and then the corresponding linear discriminant variables computed for the bootstrap sample are used as additional features for training a classification tree. The main disadvantage of double bagging is that, its success relies on the linear structure among the classes. If there is no linear relationship among the classes, adding the coefficients of LDA will result in adding some non-informative variables in the ensemble. This limitation can be overcome if the additional classifier used double bagging, has the ability learn in non-linear class structure. To overcome this in [20] authors proposed a variant of double bagging, where authors used SVM as the additional classifier model. The basic reason to use adopt SVM as the additional classifier model was for its ability to learn from non-linear space and presenting it in a linear space (with the use of suitable kernel). In that paper authors used Gaussian kernel inside the SVM and presented the performance of Double SVMBagging (which can be abbreviated as Double SVM Bagging); from now on we shall define it as decision forest with SVM as additional classifier (DF-SVM). The success of the ensemble over other ensemble methods such as, Bagging, Adaboost, Random Forest and Rotation Forest in the experiments were interesting.

In the above ensemble methods the base classifiers are trained on the bootstrap sample with the same size as the original training set, but in lot of data mining application the datasets are too large to fit in the typical computer memory. One possible approach is to use sub-sample [13] of the data. The performance of decision forest was also checked with different subsampling ratios (SSR) in [21,22]. In [21] it was shown that the with SSR = 0.50 the ensemble has better performance than other well known ensembles, later in [22] it was shown that with SSR = 0.20, 0.30 and 0.40 the performance is more better. So the evidence

is there to propose an automated sample size selection technique by which we can select the optimal subsampling ratio for the decision forest with SVM as additional classifier. In this paper we have proposed to estimate the optimal sample size for the DF-SVM with an embedded cross-validation technique.

The rest of the paper is organized as follows: The paper is organized as: in Section 2, we have described about the effect of SSR on double subagging, the subsampled version of double bagging. We have also discussed about the embedded cross-validation method in that section. In Section 3, we have stated the aim and setup of the experiments; the discussion of the results is also included in that section; this is followed by the conclusion in Section 4.

2 Decision Forest with SVM as Additional Classifier (DF-SVM)

In this section we briefly discuss about DF-SVM. The effect of small subsamples on DF-SVM is illustrated theoretically. We have also described the embedded cross-validation method.

When a decision tree is adopted as the base learning algorithm, only splits that are parallel to the feature axes are taken into account even though the decision tree is nonparametric and can be quickly trained. Considering that other general splits such as linear ones may produce more accurate trees, a "Double Bagging" method was proposed by Hothorn and Lausen [11] to construct ensemble classifiers. In double bagging framework the out-of-bag sample is used to train an additional classifier model to integrate the outputs with the base learning model. So we see that performance of the double bagging solely depends on two factors: 1) the classes of the dataset are linearly separable so that the additional predictors are informative (or discriminative), this implicitly implies that, the additional classifier should be able to discriminate the classes of the data, 2) the size of the out-of-bag samples as to construct an additional classifier model; this in turn means that the sample size of the ensemble is pivotal in better training of the additional classifier of the ensemble.

In this decision forest for each bag of instances the corresponding OOBS is utilized to construct an SVM and then the this SVM is applied back to the resampled training set to extract the CPP (Class posteriori probabilities), then all the CPPs of these SVMs are stored in the matrix CPP and are used as the additional features with the features $\mathcal{X}$ (of the resampled training set) as $[\mathcal{X}\ CPP]$ which is an $r \times p(c+1)$ matrix, where p is the number of features, r is the number of instances in the resampled training set and c is number of classes in the original training sample T. Then a base decision tree classifier is trained on this enlarged feature space. The generic framework of DF-SVM algorithm is showed in Fig 1.

2.1 Effect of Subsampling Ratios on Bagging Type Ensembles

Subsampling and m-out-of-n bootstrapping with $m < n$ originated in the statistical literature as an alternative to the bootstrap sampling [13]. The motivation

Input:

- X: Training set of size N.
- ρ: Subsample ratio. If $\rho = 1$. it will generate bootstrap samples.
- C: A base classifier.
- C^{add}: An additional classifier model, here SVM.
- B: Number of classifiers to construct.
- x: Test instance for classification.

Output: ω: Class label for x.

Step 1 **Generate training samples**
Extract a training sample of size $N * \rho$ from the training set, define this as X^b.

Step 2 **Generate Transformed Additional Predictors**
Construct additional classifier model c^{add} using the out-of-bag sample $X^{-(b)}$, Transform the original predictors $X^{(b)}$ using each additional classifier model c^{add}. Denote these as $c^{add}(x^{(b)})$.

Step 3 **Bundle Original and Additional Predictors**
Construct the combined classifier C^{comb}, by *bundling* the original predictors and the transformed predictors, as $C^{comb(b)} = (x^{(b)}, c^{add}(x^{(b)})$.
Depending on how many base classifiers we want to construct, iterate steps (2) and (3); for example B bootstrap samples.

Step 4 **Classification**
A new observation x_{new} is classified by, "majority" vote rule using the predictions of the combined B classifiers $C((x_{new}, c^{add}(x_{new})), C^{comb(b)})$ for $b = 1, 2, ...B$.

Fig. 1. Generic Framework of DF-SVM Algorithm

of these studies was to improve the efficiency of the estimates obtained by sampling with replacement. In automatic induction of models, subsampling has been studied along with bootstrapping from analytical point of view in [6,9], its different types of implementation is studied in [5], improved generalization of bagging using sample size different from original training set is studied in [16,17,23].

In [5] Breiman proposed random subsampling technique, the first ever variant of bagging that utilized sampling ratio different from the usual standard value, named *Rvotes*. Rvotes was for classification in large databases, using decision trees as base learners. In [6] *Subagging* was proposed as a computationally feasible variant of bagging. It is conjectured to have similar accuracy as bagging in regression and classification problems. The focus of these works were to setup a theoretical frame work to understand the variance reduction effect of bagging and subagging. In [6] instead of with replacement sampling a m-out-of-n without replacement sampling was used, and authors defined it as *subagging*. A suitable choice of sample size for subagging is $m_{wor} = n/2$ [9]. There is a reason to believe that the performance of the $m(= 0.5n)$-out-of-n (subsampling) bootstrapping to perform similar to n-out-of-n bootstrap. The effective size of resample in the n-out-of-n bootstrapping is n, in terms of amount of information it contains is

$$\frac{(\sum N_i)^2}{\sum N_i^2} \approx \frac{1}{2}n$$

where N_i denotes how many times the ith data value is repeated in the subsample. Following this, in [16] it was shown that the generalization performance of bagging ensemble can be improved with only 30% of the instances used for

training the bagged decision tree. In [17] authors investigated the bias-variance performance of the subsampled bagging with SVM as base classifier and showed that subsampled bagging, reduces the variance component of the error. In [23] authors reported the effect of small subsampling ratio on ensemble of stable (linear) classifiers. In that paper authors showed that the performance of subagging of stable (linear) classifiers are approximately linearly related with the subsample size, i.e., with larger subsample size the accuracy of the ensemble of stable classifiers also increase.

2.2 DF-SVM with Embedded Cross-Validation Technique

In DF-SVM ensemble method the only hyper parameter is the subsampling ratio (SSR). This parameter is pivotal in controlling the performance of the ensemble [22]. So precision is needed in selecting the optimum SSR to produce maximal performance of the DF-SVM algorithm.

In parameter selection of any predictive model, the parameter which extracts optimal generalization performance of the model is selected as the optimum parameter. In doing so it prevalent to use a validation set. The simplest case can be that, there are N models and the "best" is chosen based on the error they make on a fixed validation set of size V. Theoretically, if VE_i is the validation error of model i and TE_i is its true test error, then for all N models simultaneously the following bound holds with probability $1 - \eta$ [18]

$$TE_i \leq VE_i + \sqrt{\frac{logN - log\frac{\eta}{4}}{V}}$$

The optimality of this selection procedure depends on the number of models N and on the size V of the validation set. The bound suggests that a validation set can be used to accurately estimate the generalization performance of a relatively small number of models. This criterion can be specifically useful for selecting the size of the ensemble models, because in ensemble the computational complexity is higher than single models, so the scope for validating an ensemble model with independent sample is limited. In this paper for final selection of the subsample ratio of the DF-SVM ensembles by embedding a cross-validation process after the training of the ensembles.

To select the DF-SVM ensemble with optimum SSR we need to check the validation performance of the DF-SVM with that particular SSR. But checking the generalization performance of DF-SVM with each SSR is cumbersome. In bagging type ensembles the validation set is built in, which is constituted by the out-of-bag samples (OOBS). But in this ensemble the SVMs are trained on OOBS, so the facility of the built in test set is not there. Moreover small validation set can cause overfitting, so this should be large enough for efficient selection. To maximize the amount of available data for validation, we split the training set into equal parts and perform the validation based on each part. The pseudo code of the algorithm is given in Figure 2.

□ **Input parameter**: ρ = Number of subsampling ratios.

- For $m = 1$ to M (Number of subsampling ratios)
  1. Split the training set into V equal partitions (In this paper it is 5).

     For $v = 1$ to V
     a. Validation set $(Val) = n_{val}$ objects. The partition with number v.
     b. Training set $(Train)$ = other partitions than v.
     c. Construct the DF-SVM with ρ_m, define this model as DF-SVM$_{\rho_m}$.
     d. Apply the DF-SVM$_{\rho_m}$ to the Val set and compute the misclassification error e_{v_m} for the partition v.

  2. Compute the average validation error of DF-SVM$_{\rho_m}$ as $\mathrm{E}_{\text{DF-SVM}_{\rho_m}} = \sum_{v=1}^{V} e_{v_m}$

□ **Output parameter**: ρ_* ; for which $\mathrm{E}_{\text{DF-SVM}_{\rho_*}} = \min(\mathrm{E}_{\text{DF-SVM}_{\rho_m}})\ \forall m = 1, \cdots, M$.

Fig. 2. Pseudo code of embedded cross-validation technique used in DF-SVM algorithm

In this embedded cross-validation technique, a cross-validated model is created by training a model multiple times in different folds with the *same model parameters*, here SSR. To make a prediction for a test point, a cross-validated model simply averages the predictions made by each of the models in each repetitions. The prediction for a training point (that subsequently will be used for ensemble validation), however, only comes from the individual model that did not see the point during training. In essence, the cross-validated model delegates the prediction responsibility for a point that will be used for validation to the one sibling model that is not biased for that point. In this way we can compute the error of the embedded cross-validation method as a function of the sample size (or SSR) and compare with the test error (computed on the totally *unseen* test set) and check whether this embedded cross-validation error is 'near honest' estimator of true error.

3 Experiments and Discussion of Results

In this section we have firstly stated the aim and setup of the experiments and then we have discussed the results obtained from the experiments.

3.1 Aim and Setup of the Experiments

In this paper we have proposed an embedded cross-validation method to select the optimum sample size of DF-SVM ensembles. We conducted two sets of experiments to check the performance of the selection method on the respective ensemble. In the first experiment we compare the embedded cross-validation error of the DF-SVM with each SSR with the test error. The purpose of this experiment is to check whether the error of embedded cross-validation method can honestly represent the true error. If it does follow the test error than we can use this embedded cross-validation technique to select the optimum SSR for

DF-SVM. In the second experiment we have compared the optimal DF-SVM (opDF-SVM) with several most popular ensemble methods, bagging, adaboost and rotation forest. In all our experiments we have used 15 UCI [1] benchmark datasets. The dataset descriptions are given Table 1. In the second experiment we have reported average error of 20 three-fold cross-validation results i.e., each cell in the Table 2 consists of an average value of total 60 (20×3-CV) testing.

Table 1. Description of the 15 Data used in this paper

Dataset	Objects	Classes	Features
Australian Credit	690	2	16
Balance Scale	625	3	5
Breast Cancer	286	2	9
Diabetes	768	2	8
Ecoli	336	8	8
German-credit	1000	2	20
Glass	214	7	9
Cleveland-Heart	297	5	13
Heart-Statlog	270	2	13
Ionosphere	351	2	34
Iris	150	3	4
Japanese Credit	690	2	15
Liver-disorder	345	2	6
Lymphography	148	4	18

For all the ensemble methods we have used a full decision tree as the base classifier. The size of all ensembles is fixed to 50. We have used subsample ratios $0.1 - 0.8$ to build the DF-SVM ensembles. For the second experiment we performed the comparison of the methods as proposed by Demšar [7], which is described below:

- First perform the Friedman test, to check the null hypothesis that all methods have equal performance.
- Then if the null hypothesis is rejected, perform the Nemenyi posthoc test to detect the methods which are not significantly different from each other.
- Perform the Wilcoxon Sign Rank test, where we compare the *best* performing classifier with all other classifiers. We select best classifier on the basis of lowest average rank for errors. If there are several classifiers with lower average rank for errors with very small difference, then we select the one with lowest average rank for training time among them.

We have notified the classifiers of the same group as, 1 or 2 depending on how many groups they can divided; here the groups are sorted in ascending order, so that no. 1 will correspond as the best group and so on. The Wilcoxon sign rank test will clarify the performance of the *best* ensembles in Table 2 if there is no significant difference detected by the Nemenyi posthoc test.

3.2 Discussion of Results

We have presented the results of the first experiment in Figure 3. For lack space we have presented here the results of 8 datasets. We see that for the dataset the embedded cross-validation error nicely corresponding with the test error. This implies that we can use the embedded cross-validation technique to select the optimum SSR for DF-SVM. It is very easy to note that the performance of DF-SVM is far better within SSR: 0.20 – 0.50 than with higher SSR. So in the final experiment we will confine our selection only with SSR = 0.20 – 0.50.

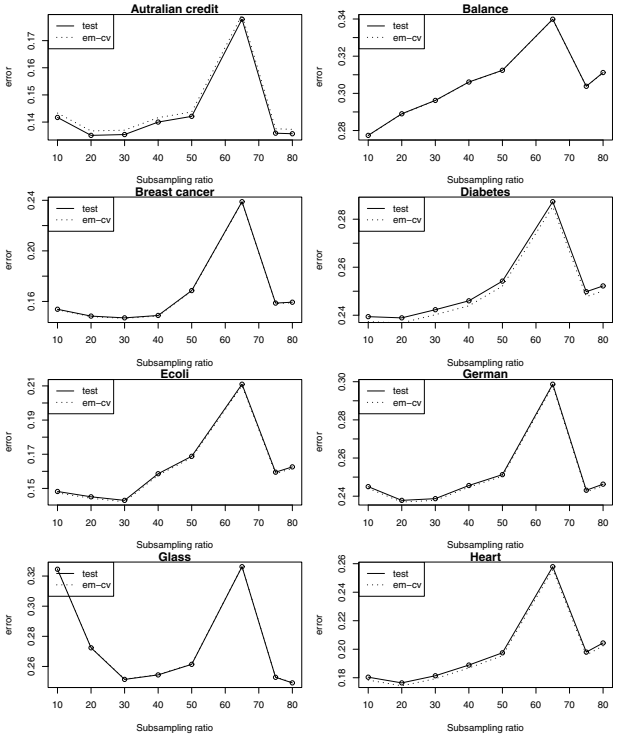


Fig. 3. The embedded cross-validation and test error as a function of the subsampling ratio of DF-SVM

In Table 2 we have presented all the misclassification error opDF-SVM, Bagging, Adaboost, Multiboost and Rotation Forest. We can se that the performance of the opDF-SVM is surprisingly better than all the ensemble methods in most of the datasets. We have grouped the algorithms according to the Friedman-Nemenyi test (defining the groups as F-N Groups) in htm last row of the table and see that the group opDF-SVM is in there is no other algorithm and also it is with the lowest average rank (having the highest accuracy).

Table 2. Error of opDF-SVM Bagging, AdaBoost, Multiboost and Rotation Forest

Dataset	opDF-SVM	Bagging	Adaboost	Multiboost	Rotation Forest
Australian	0.1347	0.1329	0.1358	0.1322	0.1364
Bcs	0.2879	0.2944	0.3109	0.3020	0.3102
Balance	0.1546	0.1520	0.2076	0.1813	0.1621
Diabetes	0.2489	0.2504	0.2556	0.2498	0.2543
Ecoli	0.1306	0.1772	0.1560	0.1509	0.1729
German	0.2125	0.2836	0.2543	0.2496	0.2880
Glass	0.2137	0.2448	0.2461	0.2430	0.2439
Cleveland	0.1866	0.1988	0.2099	0.2076	0.1975
Hearts	0.1556	0.2368	0.2092	0.2034	0.2110
Ionosphere	0.0617	0.0876	0.0621	0.0634	0.0532
Iris	0.0526	0.0561	0.0411	0.0529	0.0582
Japanese	0.1243	0.1334	0.1249	0.1342	0.1302
Liver	0.2358	0.2983	0.3067	0.2912	0.2961
Lymph	0.1947	0.3391	0.2812	0.2269	0.3021
F-N Groups	1	3	2	2	2

4 Conclusion

The cross-validation technique is embedded inside the DF-SVM ensemble to select the optimum SSR in this paper. In the DF-SVM the facility of the built in test sample is utilized to train the SVM model, so to build up an automatic selection of sample size with honest selection criterion. The splitting of the training set to train and (then) test the models serve this purpose of test set. It is computationally feasible to select the optimum SSR in this way because it is carried out with 5-fold cross-validation. This technique is theoretically feasible also as the embedded cross-validation error of the DF-SVM conforms with test error. We see that the use of this embedded cross-validation method greatly increases the performance of the respective ensemble. The benefit is mainly due to having more data for validation, the other reason for the success is having a larger OOBS to train the additional classifier model. From our experimental results it is clear that the emebedded cross-validation technique can represent the true test error and hence can be used to as validation technique to select optimum SSR. And finally the comparison with other well known ensembles also confirm that optimum SSR selection technique can improve the performance of the DF-SVM.

References

1. Asuncion, A., Newman, D.J.: UCI Repository of Machine Learning Databases (2007), http://www.ics.uci.edu/mlearn/MLRepository.html
2. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996a)

3. Breiman, L.: Out-of-bag estimation. Statistics Department, University of Berkeley CA 94708, Technical Report (1996b)
4. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
5. Breiman, L.: Pasting small votes for classification in large databases and on-line. Machine Learn. 36(1-2), 85–103 (1999)
6. Bühlman, P.: Bagging, subagging and bragging for improving some prediction algorithms. In: Arkitas, M.G., Politis, D.N. (eds.) Recent Advances and Trends in Nonparametric Statistics, pp. 9–34 (2003)
7. Demšar, J.: Statistical comparisons of classifiers over multiple datasets. J. Mach. Learn. Research 7, 1–30 (2006)
8. Freund, Y., Schapire, R.: Experiments with a New boosting algorithm. Machine Learning. In: Proceedings to the Thirteenth International Conference, pp. 148–156. Morgan Kaufmann, San Francisco (1996)
9. Friedman, J., Hall, P.: On Bagging and Non-linear Estimation. J. Statist. Planning and Infer. 137(3), 669–683 (2007)
10. Hastie, T., Tibshirani, R., Freidman, J.: The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, New York (2009)
11. Hothorn, T., Lausen, B.: Double bagging: combining classifiers by bootstrap aggregation. Pattern Recognition 36(6), 1303–1309 (2003)
12. Kuncheva, L.I.: Combining Pattern Classifiers. Methods and Algorithms. John Wiley and Sons, Chichester (2004)
13. Politis, D., Romano, J.P., Wolf, M.: Subsampling. Series in Statistics. Springer, Berlin (1999)
14. Rodríguez, J., Kuncheva, L., Alonso, C.: Rotation forest: A new classifier ensemble method. IEEE Trans. Patt. Analys. Mach. Intell. 28(10), 1619–1630 (2006)
15. Schapire, R.: The Boosting Approach to Machine Learning: An overview. In: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, Y. (eds.) MSRI Workshop on Nonlinear Estimation and Classification. Lecture Notes in Statistics, vol. 171, pp. 149–172. Springer, Heidelberg (2002)
16. Terabe, M., Washio, T., Motoda, H.: The effect of subsampling rate on subagging performance. In: Proceedings of ECML 2001/PKDD2001 Workshop on Active Learning, Database Sampling,and Experimental Design: Views on Instance Selection, pp. 48–55 (2001)
17. Valentini, G.: Random Aggregated and Bagged Ensembles of SVMs: An Empirical Bias–Variance Analysis. In: Roli, F., Kittler, J., Windeatt, T. (eds.) MCS 2004. LNCS, vol. 3077, pp. 263–272. Springer, Heidelberg (2004)
18. Vapnik, V.: Statistical Learning Theory. Springer, New York (1999)
19. Webb, G.I.: MultiBoosting: A technique for combining boosting and wagging. Machine Learning 40(2), 159–196 (2000)
20. Zaman, F., Hirose, H.: A new double bagging via the support vector machine with application to the condition diagnosis for the electric power apparatus. Lecture Notes in Engineering and Computer Science, vol. 2174(1), pp. 654–660 (2009a)
21. Zaman, F., Hirose, H.: Double SVMBagging: A new double bagging with support vector machine. Engineering Letters 17(2), 128–140 (2009b)
22. Zaman, F., Hirose, H.: Double SVMSbagging: A subsampling approach to SVM ensemble. Intelligent Automation And Computer Engineering. Lecture Notes in Electrical Engineering, vol. 52, pp. 387–399. Springer, Heidelberg (2010)
23. Zaman, F., Hirose, H.: Effect of Subsampling Rate on Subbagging and Related Ensembles of Stable Classifiers. In: Chaudhury, S., Mitra, S., Murthy, C.A., Sastry, P.S., Pal, S.K. (eds.) PReMI 2009. LNCS, vol. 5909, pp. 44–49. Springer, Heidelberg (2009)

Combining Classifier with a Fuser Implemented as a One Layer Perceptron

Michal Wozniak and Marcin Zmyslony

Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
{Michal.Wozniak,Marcin.Zmyslony}@pwr.wroc.pl

Abstract. The combining approach to classification so-called Multiple Classifier Systems (MCSs) is nowadays one of the most promising directions in pattern recognition and gained a lot of interest through recent years. A large variety of methods that exploit the strengths of individual classifiers have been developed. The most popular methods have their origins in voting, where the decision of a common classifier is a combination of individual classifiers' outputs, i.e. class numbers or values of discriminants. Of course to improve performance and robustness of compound classifiers, different and diverse individual classifiers should be combined. This work focuses on the problem of fuser design. We present some new results of our research and propose to train a fusion block by algorithms that have their origin in neural computing. As we have shown in previous works, we can produce better results combining classifiers than by using the abstract model of fusion so-called *Oracle*. The results of our experiments are presented to confirm our previous observations.

Keywords: combining classifier, multiple classifier system, fuser design, one-layer perceptron.

1 Introduction and Related Works

The aim of a recognition task is to classify a given object by assigning it (on the basis of observing the features) to one of the predefined categories [4]. There are many propositions on how to automate the classification process. We could use a number of classifiers for each task, but according to the „no free lunch theorem" there is not a single solution that could solve all problems, but classifiers have different domains of competence [19]. Let's have a look at the sources of classifier misclassification:

- Firstly, the classifier usually uses model that does not fit into the real target concept (e.g. model is simplified because of costs),
- Learning material, i.e., learning set is limited, unrepresentative, or includes errors.

Fortunately we are not doomed to failure because for each classification task we could usually use many classifiers. The best one could be chosen on the basis of the evaluation process or we could use the whole pool of available classifiers. Let's note that usually an incompetence area, i.e. subset of feature space where all individual classifiers make a wrong decision is very small what is shown in Fig. 1.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 282–291, 2011.

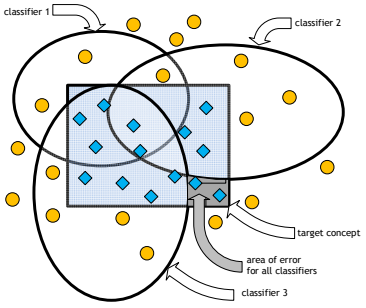


Fig. 1. Decision areas of 3 different classifiers for a toy problem

This observation explains why works devoted to multiple classifier systems (MCSs) are currently the focus of intense research. The main motivations of using MCSs are as follows:

- MCSs could avoid selection of the worst classifier for small sample [16];
- there is evidence that classifier combination can improve the performance of the best individual ones because it exploits unique classifier strengths [18];
- combined classifiers could be used in a distributed environment like distributed computing systems (P2P, GRID) [5], especially in the case of a database that is partitioned for privacy reasons and in each node of the computer network only a final decision can be available.

Let's underline that designing a MCS is similar to the design of a classical pattern recognition [10] application. When designing a typical classifier, the aim is to select the most valuable features and choose the best classification method from the set of available ones. The design of a classifier ensemble is similar – it is aimed to create a set of complementary/diverse classifiers. The design of a fuser is aimed to create a mechanism that can exploit the complementary/diversity of classifiers from an ensemble and combine them optimally.

The typical architecture of MCSs is depicted in Fig.2.

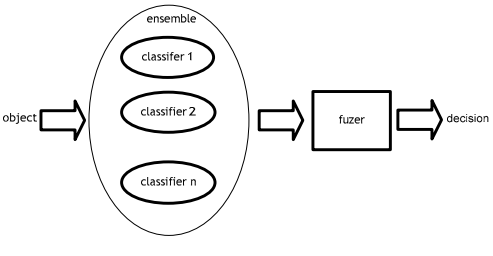


Fig. 2. Model of MCS

There are a number of important issues when building the MCSs, which can be grouped into two main problems:

- How to design the classifier ensemble,
- How to design the fuser.

Apart from increasing the computational complexity, combining similar classifiers should not contribute to the system becoming constructed. Therefore, selecting members of the committee with different components seems interesting. An ideal ensemble consists of classifiers with high accuracy and high diversity, i.e. mutually complementary. Several papers introduce different types of diversity measures that allow for the possibility of a coincidental failure to be minimized [1]. A strategy for generating the ensemble members must seek to improve the ensemble's diversity. To enforce classifier diversity we could use varying components of the MCS:

- different input data e.g., we could use different partitions of a data set or generate various data sets by data splitting, cross-validated committee, bagging, boosting [15], because we hope that classifiers trained on different inputs are complementary;
- classifiers with different outputs, i.e. each individual classifier could be trained to recognize a subset of only predefined classes (e.g., binary classifier - one class against rest ones strategy) and the fusion method should recover the whole set of classes. A well known technique is Error-Correcting Output Codes [6];
- classifiers with the same input and output, but trained on the basis of a different model or model's versions.

The problem of assuring high diversity of classifier ensemble is crucial for quality of above mentioned compound classifiers but our paper focus on the second important issue of MCS design – the problem of fuser design. The following sections will present a taxonomy on fuser design and properties of some fusers confirmed both analytically and experimentally.

2 Fuser Design

There are several propositions on how to negotiate a common decision by the group of classifiers and having a choice of a collective decision making method is an important issue of MCS design. The first group of methods includes algorithms using discrete outputs (class numbers) of individual classifiers only [17]. It is a quite intuitive and flexible proposition because it can combine outputs of classifiers using different pattern recognition models. Initially only majority-voting schemes were implemented, but in later work more advanced methods were proposed.

Many known conclusions regarding the classification quality of MCSs have been derived analytically, but are typically valid only under strong restrictions, such as particular cases of the majority vote [11] or make convenient assumptions, such as the assumption that the classifier committee is formed from independent classifiers. Unfortunately, such assumptions and restrictions are of a theoretical character and not

useful in practice. From aforementioned research we can make the following conclusion that it is worthy to combine classifiers only if the difference among their qualities is relatively small. It has to be noted that the higher the probability of misclassification of the best classifier, the smaller quality difference should be in order to get an effective committee that outperforms their components. Some additional information about voting classifier quality can be found in [1, 15].

For this kind of fusion the Oracle classifier is usually used as a reference combination method. Many works consider the quality of the Oracle as the limit of the quality of different fusion methods [20]. It is an abstract fusion model, where if at least one of the classifiers recognizes an object correctly, then it points at the correct class too. The Oracle is usually used in comparative experiments to show the limits of classifier committee quality [17].

2.1 Fuser Based on Classifier Response

The formal model of fusion based on classifier responses is as follows. Let us assume that we have n classifiers $\Psi^{(1)}, \Psi^{(2)}, ..., \Psi^{(n)}$ and each of them decides if a given object belongs to class $i \in \boldsymbol{M} = \{1, ..., M\}$. The decision rule of combining classifier $\overline{\Psi}$ is as follows :

$$\overline{\Psi} = \arg\max_{j \in \boldsymbol{M}} \sum_{l=1}^{n} \delta\left(j, \Psi^{(l)}\right) w^{(l)} \Psi^{(l)} \tag{1}$$

where $w^{(l)}$ is the weight of the l-th classifier and

$$\delta(j,i) = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases}. \tag{2}$$

Let us note that $w^{(l)}$ plays a key-role of the quality of the classifier $\Psi^{(l)}$. Much research has been done on how to set the weights e.g., in [21] authors proposed to learn the fuser. Let us consider three possible weight set-ups:

- weights dependent on the classifier,
- weights dependent on the classifier and the class number,
- weights dependent on the features value, the classifier, and the class number.

2.2 Fuser Based on Discriminant

Let us consider an alternative model for the construction of a combining classifier, one that performs classifier fusion on the basis of the discriminants of individual classifiers. It is a less flexible model than presented in the previous section because some restrictions are placed on a group of individual classifiers. Firstly, they should make decisions on so-called discriminants (support function). The main form of discriminants are posterior probability estimators, typically associated with probabilistic models of the pattern recognition task [4], but it could be given for e.g. by the output of neural networks or that of any other function whose values are used to establish the decision of the classifier.

One concept is known as the Borda count. A classifier based on this concept makes decisions by giving each class support corresponding to the position in the ranking. For this form of fusion we could use classifiers based on different models because class ranks are required only. It is worth noting that such methods of fusion could outperform the Oracle model.

The remaining methods of fusion based on discriminants place another restriction on the classifier ensemble. They require that all classifiers in the committee use the same pattern recognition model i.e., that mathematical interpretation of the support function should be the same for each of them. The aggregating methods, which do not require learning perform fusion with the help of simple operators, such as the maximum, minimum, average, or product are typically relevant only in specific, clearly defined conditions [7]. Weighting methods are an alternative and the selection of weights has a similar importance as in the case of weighted voting [3, 13].

The formal model of fusion based on discriminants is as follows. Let us assume that we have n classifiers $\Psi^{(1)}, \Psi^{(2)}, ..., \Psi^{(n)}$ and each of them makes a decision on the basis of the values of discriminants. Let $F^{(l)}(i, x)$ denotes a function that is assigned to class i for a given value of x, and which is used by the l-th classifier $\Psi^{(l)}$. A common classifier $\hat{\Psi}(x)$ is described as follows [12]

$$\hat{\Psi}(x) = i \quad \text{if} \quad \hat{F}(i, x) = \max_{k \in M} \hat{F}(k, x), \tag{3}$$

where

$$\hat{F}(i, x) = \sum_{l=1}^{n} w^{(l,i)} F^{(l)}(i, x) \text{ and } \sum_{i=1}^{n} w^{(l,i)} = 1 . \tag{4}$$

Let us consider four possible weight set-ups for a fuser that is based on the values of the classifiers' discriminates:

- weights dependent on the classifier,
- weights dependent on the classifier and the feature vector,
- weights dependent on the classifier and the class number,
- weights dependent on the classifier, the class number, and the feature vector.

If we consider the two class recognition problem only for the last two cases where weights are dependent on classifier and class number it is possible to produce a compound classifier that could achieve quality equal to or better than the Oracle [23]. But when we take into consideration more than two class recognition problems we could observe that it is possible in all aforementioned cases to get results better than the Oracle. Weights independent from x could be assigned to a linear separated problem, in other cases we should use weights that depend on classifier, class number, and feature vector values. More details and an illustrative example of the features of weighted voting using a weight dependent classifier and class numbers can be found in [22].

3 Fuser Learning Algorithm

In this paper we look at the case where weights are dependent on the classifier and the class number, because cases where weights are dependent on the feature vector are de facto function estimation problems that require additional assumptions about the mentioned above functions and usually lead to a parametric case of function estimation. The considered case does not require additional assumptions and the formulation of the optimization task is a quite simple.

Let us present an optimization problem that will return minimal misclassification results for the fusion of the classifier.

3.1 Optimization Problem

For the case where weights are dependent on the classifier and the class number fuser learning task leads to the problem of how to establish the following vector W

$$W = \left[W^{(1)}, W^{(2)}, ..., W^{(n)} \right] \tag{5}$$

which consists of weights assigned to each classifier (denoted as l) and each class number.

$$W^{(l)} = \left[w^{(l)}(1), w^{(l)}(2), \ldots, w^{(l)}(M) \right]^T \tag{6}$$

We could formulate the following optimization problem. The weights should be established in such a way as to maximize the accuracy probability of the fuser:

$$\Phi(W) = 1 - P_e(W), \tag{7}$$

where $P_e(W)$ is probability of misclassification.

In order to solve the aforementioned optimization task, we could use one of a variety of widely used algorithms. In this work we have decided to engage in neural networks.

3.2 Neural Algorithm

Neural networks project complex relationships between inputs and outputs and the aim of their training procedure is to solve optimization problems. In our case we decided to use a neural network as a fuser where input values get discriminants. The picture below presents the model of a one layer neural network applied to the problem under consideration.

4 Experiments

4.1 Set Up of Optimization Task

The aim of the experiments is to evaluate the performance of the fuser based on weights dependent on the classifier and the class number.

All experiments were carried out in the *Matlab* environment using dedicated software called *PRTools* [8] along with our own software.

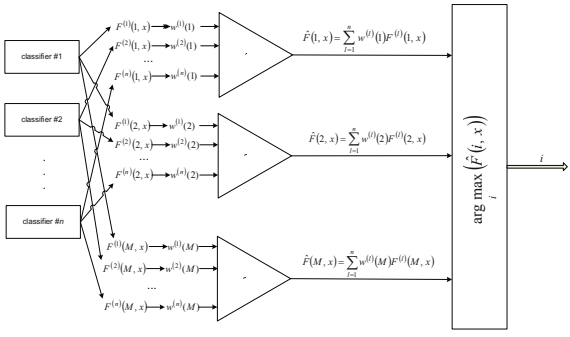


Fig. 3. Weights dependent on classifier and class number

For the purpose of this experiment, there were five neural networks prepared that could be treated as individual classifiers. The details of the used neural nets are as follows:

- five neurons in the hidden layer,
- sigmoidal transfer function,
- back propagation learning algorithm,
- number of neurons in last layer equals number of classes of given experiment.

To ensure the diversity of the simple classifiers, all were slightly undertrained (the training process was stopped early for each classifier).

To evaluate the experiment we used ten databases from the UCI Machine Learning Repository [2], which are described in Tab. 1.

Table 1. Databases' description

	Dataset	Number of (1) attributes, (2) classes, (3) examples		
		(1)	(2)	(3)
1	Balance_scale	4	3	625
2	Breast Cancer Wisconsin	10	2	699
3	Connectionist Bench (Vowel Recognition)	10	11	528
4	Glass_Identification	9	7	214
5	Haberman	3	2	306
6	Image_segmentation	19	7	2100
7	Ionosphere	34	2	351
8	Letter_Recognition	16	26	20000
9	MAGIC Gamma Telescope	10	2	17117
10	Yeast	9	10	1484

For each database the experiment was repeated ten times with a different epoch of learning. The best results obtained in those experiments are presented in the table below with additional information about the result obtained by the *Oracle* classifier.Statistical differences between the performances of the classifiers were evaluated using 10-Fold Cross-Validated Paired t Test. The results are presented in Tab. 2.

Table 2. Classification errors

	Dataset	Oracle	NN-fuzer	MV
1	Balance_scale	6.76%	8.36%	14.94%
2	Breast Cancer Wisconsin	3.09%	1.95%	38.55%
3	Connectionist	21.23%	17.78%	79.51%
4	Glass_Identification	39.79%	31.94%	69.11%
5	Haberman	19.30%	26.20%	51.09%
6	Image_segmentation	29.47%	9.89%	38.25%
7	Ionosphere	8.57%	7.94%	14.92%
8	Letter_Recognition	89.00%	95.08%	95.14%
9	MAGIC Gamma Telescope	13.82 %	15.70%	35.30%
10	Yeast	1.18%	15.91%	25.48%

We can state, that according to 10-Fold Cross-Validated Paired t Test in eight cases (1, 2, 3, 4, 7, 8, 9, 10) we can confirm our hypothesis that the proposed fuser outperforms the Oracle classifier.

4.2 Results Evaluation

The results presented in Tab. 2 prove that neural networks are efficient tools for solving optimization problems. As stated before, when weights are dependent on the classifier and the class number, it is possible to achieve results that are better than the *Oracle* classifier. Unfortunately we can not formulate a general conclusion on the basis of the experiments that were carried out because we still do not know what conditions should be fulfilled to produce a high quality combining classifier used in the proposed fusion method. We would like to underline that we observed that a fuser based on weights dependent on the classifier and the class number could outperform Oracle classifier but it does not guarantee it. As we mentioned above we still have not discovered conditions that should be fulfilled to produce the desirable fuser. It probably depends on the conditional probability distributions of classes for the given classification problem what was partially confirmed by our analytical research [23].

We should always remember that the tools that were used to solve the optimization task in our experiments are somehow black box and only the appropriate settings of all the parameters can return satisfactory results.

5 Final Remarks

The method of classifier fusion that uses weights dependent on the classifier and the class number was discussed in this paper. Additionally the method of fuser training based on the neural computing approach was presented and evaluated via computer experiments carried out on several benchmark datasets.

The results obtained justify the use of the weighted combination and they are similar to what was published in [13]. Unfortunately, as previously stated, it is not possible to determine the values of the weights in an analytical way, therefore using heuristic methods of optimization (like neural or evolutionary algorithms) seem to be a promising research direction.

Acknowledgments. This research is supported by The Polish Ministry of Science and Higher Education under the grant which is realizing in years 2010-13.

References

1. Alexandre, L.A., Campilho, A.C., Kamel, M.: Combining Independent and Unbiased Classifiers Using Weighted Average. In: Proc. of the 15th Internat. Conf. on Pattern Recognition, vol. 2, pp. 495–498 (2000)
2. Asuncion, A., Newman, D.J.: UCI ML Repository, School of Information and Computer Science. University of California, Irvine (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
3. Biggio, B., Fumera, G., Roli, F.: Bayesian Analysis of Linear Combiners. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 292–301. Springer, Heidelberg (2007)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
5. Chmaj, G., Walkowiak, K.: Preliminary study on optimization of data distribution in resource sharing systems. In: Proc. of the 19th Internat. Conf. on Systems Engineering, ICSEng 2008, pp. 276–281 (2008)
6. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)
7. Duin, R.P.W.: The Combining Classifier: to Train or Not to Train? In: Proc. of the ICPR2002, Quebec City (2002)
8. Duin, R.P.W., et al.: PRTools4, A Matlab Toolbox for Pattern Recognition, Delft University of Technology (2004)
9. Fumera, G., Roli, F.: A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems. IEEE Trans. on PAMI 27(6), 942–956 (2005)
10. Giacinto, G.: Design Multiple Classifier Systems, PhD thesis, Universita Degli Studi di Salerno (1998)
11. Hansen, L.K., Salamon, P.: Neural Networks Ensembles. IEEE Trans. on PAMI 12(10), 993–1001 (1990)
12. Jackobs, R.A.: Methods for combining experts' probability assessment. Neural Computation 7, 867–888 (1995)
13. Jackowski, K.: Multiple classifier system with radial basis weight function. In: Graña Romay, M., Corchado, E., Garcia Sebastian, M.T. (eds.) HAIS 2010. LNCS, vol. 6076, pp. 540–547. Springer, Heidelberg (2010)

14. Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition 34, 299–314 (2001)
15. Kuncheva, L.I.: Combining pattern classifiers: Methods and algorithms. Wiley, Chichester (2004)
16. Marcialis, G.L., Roli, F.: Fusion of Face Recognition Algorithms for Video-Based Surveillance Systems. In: Foresti, G.L., Regazzoni, C., Varshney, P. (eds.) Multisensor Surveillance Systems: The Fusion Perspective, Kluwer Academic Publishers, Dordrecht (2003)
17. Tumer, K., Ghosh, J.: Analysis of Decision Boundaries in Linearly Combined Neural Classifiers. Pattern Recognition 29, 341–348 (1996)
18. Van Erp, M., Vuurpijl, L.G., Schomaker, L.R.B.: An overview and comparison of voting methods for pattern recognition. In: Proc. of IWFHR. 8, Canada, pp. 195–200 (2002)
19. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications (2001)
20. Woods, K., Kegelmeyer, W.P.: Combination of multiple classifiers using local accuracy estimates. IEEE Transactions on PAMI 19(4), 405–410 (1997)
21. Wozniak, M., Jackowski, K.: Some remarks on chosen methods of classifier fusion based on weighted voting. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS, vol. 5572, pp. 541–548. Springer, Heidelberg (2009)
22. Wozniak, M., Zmyslony, M.: Fuser on the basis of discriminants evolutionary and neural methods of training. In: Corchado, E., Graña Romay, M., Manhaes Savio, A. (eds.) HAIS 2010. LNCS, vol. 6077, pp. 590–597. Springer, Heidelberg (2010)
23. Wozniak, M., Zmyslony, M.: Fusion methods for the two class recognition problem - analytical and experimental results. In: Ryszard, Choraś, S. (eds.) Image processing and communications challenges 2, pp. 135–142. Springer, Heidelberg (2010)

Search Result Clustering Using Semantic Web Data

Marek Kopel and Aleksander Zgrzywa

Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
{marek.kopel,aleksander.zgrzywa}@pwr.wroc.pl
http://www.zsi.pwr.wroc.pl/~{kopel,zgrzywa}

Abstract. Traditional Web (Web 1.0) is a web of documents. Finding documents is the main goal of information retrieval. There were some improvements in IR (Information Retrieval) on the Web since *tf-idf* (term frequencyinverse document frequency) concerning using other information than just documents themselves. One of those approaches is analyzing link structure used in HITS and Google PageRank. Another approach may be using time metadata to enable filtering based on document publishing date as used e.g. in Google Blog Search. In this paper a Web IR method using relationship metadata and clustering is presented.

Keywords: clustering, IR, reranking, Semantic Web.

1 Relationships on the Web

Semantic Web envisioned in [1] is a web of data, data that allow a machine "understand" information on the original Web; data that express how the information relates to real world objects. The semantic data or metadata may also concern the relationships between objects. Since early versions of HTML the <a> tag used for defining hyperlinks provided attributes rel and rev meant for expressing type of relationship between the two linked objects. The attributes could be used not only for relationships between documents (e.g. table of content, chapter, subsection, index, glossary). Using *rev="made"* could be used to identify the document author. A link using attribute *rev* value could point to either the author's email address with a *mailto* URI to the author's home page. In 2004, inspired by this notation, a microformat called *rel-tag* was introduced. Using attribute value *rel="tag"* allow content tagging, i.e. describing a relationship between a document and a term (keyword) or even a concept. Another microformat - XFN (XHTML Friends Network) - can be used for expressing social relationships. XFN is mostly used in blogrolls by linking from a blog to other blogs authored by friends of the blog's author.

The above relationships can be used for a personalized search and to improve accuracy of generic Web search results. To use the metadata on the relationships a model is needed.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 292–301, 2011.

2 DAC Graph

We propose DAC graph to model Web objects and their relationships. Using a graph as a model is intuitive, since most of Web algorithms like HITS described in [2] or PageRank, initially presented in [3], modeled Web as a graph. But their models only considered one type of Web objects, i.e. documents. As described in Definition 1, DAC uses three types of Web objects.

Definition 1. *A DAC graph is a pair (V,E) of sets such that V is a set of vertices and E is a set of edges. Each element of set V must be of one of types: D - document, A - author or C - concept. Each element of set E is a 2-element subset of V. Set E must not contain edges at the same time related to a vertex of type A and a vertex of type C.*

Process of constructing a DAC graph is presented in Fig. 1. Document objects are obtained by performing a Web search. In return a document collection is formed, which is a base for extracting authors and concepts described by tags. From corresponding Semantic Web documents author relationships and concept relationships are extracted. Document relationships are obtained partly from semantic data and partly from the collection itself e.g. from hyperlink references.

The concept of DAC was previously published in [4] and some early ideas can be found in [5].

3 DAC Clustering and Search Results Reranking

The main idea for using DAC for improving the accuracy of Web search results is clustering. Since the clustering is meant to be performed at each Web search, a quick clustering algorithm was needed. For test purposes an MCl algorithm introduced in [6] was chosen. The clusters of DAC graph are used for creating a better ranking than the one of results from search performed for DAC creation. Since the new ranking is using ranks from the original one, we called it reranking. The main concept of the algorithm can be described by the following pseudo-code:

```
rerankSearchResults() {
  cluster(DAC, using MCl);
  FOR EACH (cluster IN clusters) {
    ORDER(documents in cluster, BY original ranks, ASCending);
    calculate(clusterRank, using Formula 1);
  }
  ORDER(clusters, BY clusterRank, ASCending);
}
```

$$R(C_i) = \overline{r(C_i)} - \frac{\sigma(r(C_i))}{max(\sigma(r(C_i)))} \quad . \tag{1}$$

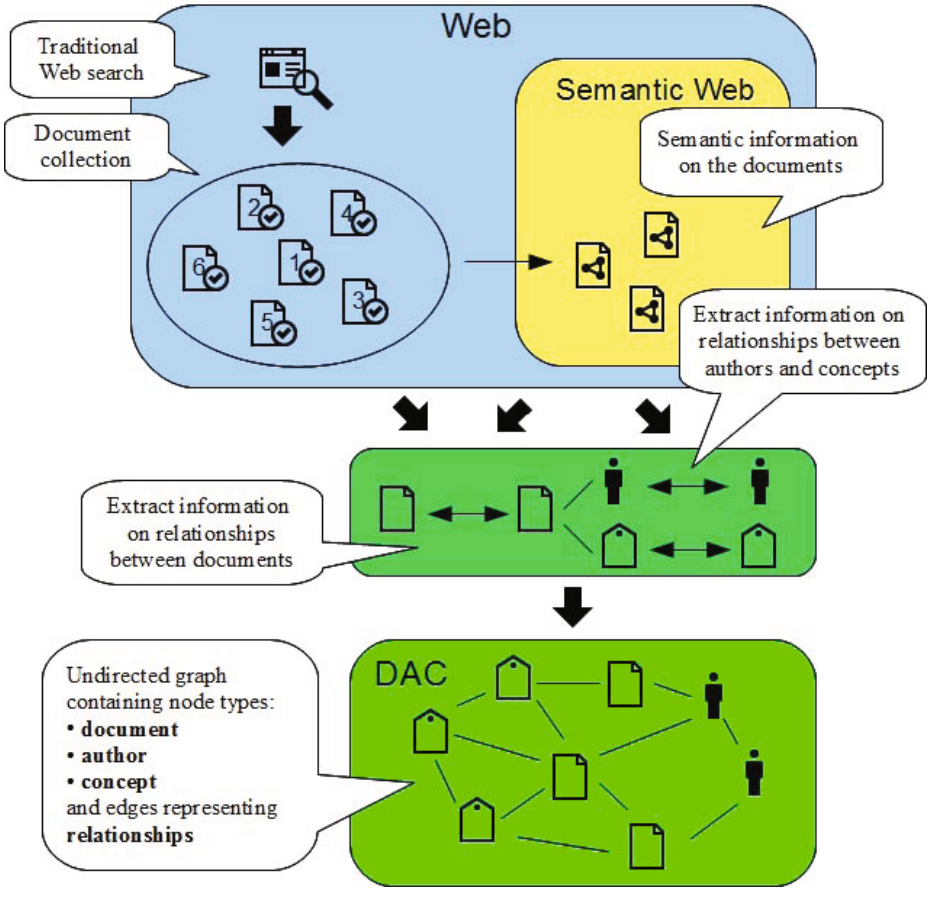


Fig. 1. Algorithm schema for constructing DAC graph. Documents vertices for the graph are given by a search result in Web 1.0. Semantic Web, which is a part of the Web contain RDF statements corresponding to the retrieved documents. From these statement author and concept vertices are extracted, as well as their relationship information. The relationships are modeled as DAC graph edges.

The intuition behind the rank calculated by Formula 1 is that a cluster (Ci) is relatively (the overline means the average) the higher in the ranking (smaller rank) the bigger the standard deviation (sigma) of the original ranks (lowercase "r") gets. This way documents with poor original ranks, but highly related to the documents with high ranks, are presented right below them.

4 Test Data and DAC Clustering Search Engine

In order to verify the reranking algorithm some real test data were needed. It is not easy to find a Web site offering documents with semantic information on their authors and concepts used, not to mention their relationships. However it turned out that using an RSS channel of blogging platform WordPress.com, their blogroll and comments plus an external Wordnet::Similarity module, can assure bringing all the needed data.

To automate the data processing and for further reranking evaluation a DAC clustering search engine was build. This way all the objects and relationships could have been indexed and estimated in the pre-search phase. Having the data prepared allow instant construction of a DAC graph instance and real time clustering for creating a reranking each time user submits a query.

First thing, to index Web object for DAC, was to parse RSS channels of WordPress blogs. It allowed extracting posts (documents) with corresponding authors and tags (concepts). Out of 26 000 WordPress blogs 129 000 post for a selected month have been indexed. Then parsing RSS for each post's comments, thanks to Pingback support, the direct link element of document relationships could have been extracted. The hyperlink connection and a normalized difference between two document original ranks are used to calculate the weight of the DAC edge incident with the two document vertices. The document-author/concept edge weight is calculated as a normalized inverse of the number of authors/concepts of the document.

The author-author relationships are scraped from blogroll XFN links. The links create a social graph of 2362 blog authors and 3462 XFN relationships as presented on Fig. 2. It shows that blogrolls are not as popular as one would want. Since this social graph is rather sparse (number of edges make 0.124% of number of edges in a complete graph with he same number of vertices), all the authored blog vertices are already clustered into 703 disconnected subgraphs. Only a few subgraphs contain more than 15 verices. Nevertheless this social graph is used to calculate weights for edges connecting author vertices in DAC graph. Each weigh is the inverse of the length of shortest path between two authors in the social graph.

For calculating relationships between concepts a Wordnet ontology is used. Using hypernym ("is a") relationships a hierarchy graph for concepts can be built. The weighs for DAC edges connecting concepts are calculated analogically to those connecting authors: it is the inverse of the number of edged between the concepts in the Wordnet hierarchy. The selected month posts are described with more than 400 000 different tags. For calculating the relationship value (edge

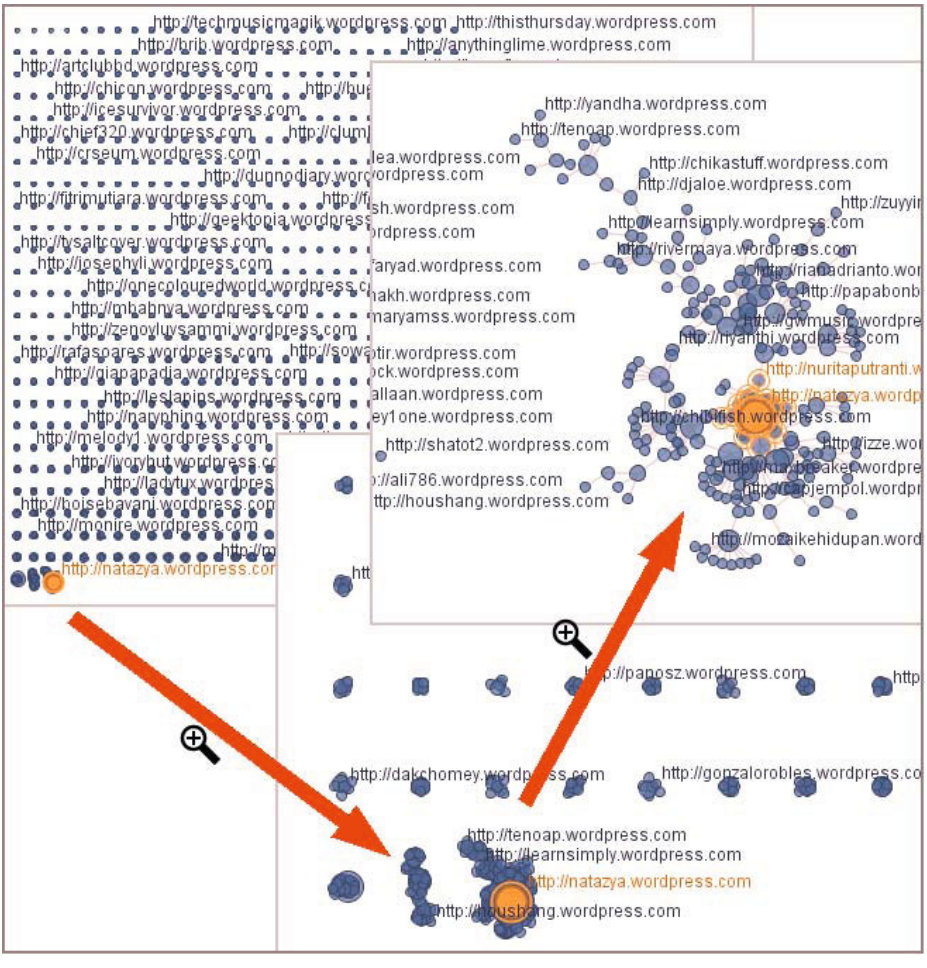


Fig. 2. Social graph of wordpress.com blog users created by HTML scraping of blogroll XFN (XHTML Friends Network) hyperlinks. Visualization created using ManyEyes (IBM).

weight) for each pair the number of tags needed some optimization. First, after converting to lower case and filtering using "^[a-z]*$" regular expression (tags only containing small letters and spaces), we get 154 000 of tags. Then using stemming and selecting only those tags that exist in Wordnet we get 21 000. Eventually, further narrowing down the tags to only those that are used more than a few times, we get 2 000 concepts.

5 Verification

To verify whether DAC reranking improves user experience, user feedback on result relevance is needed. Users were asked to mark pertinent documents for each query result. "Pertinent" means "relevant to user needs, not necessarily relevant to query". Having the pertinent documents marked once for a query we can compare their order in the original ranking - based only on term weighting - and the reranking based on DAC clustering.

5.1 Measures

Traditional measures for IR are precision and recall. However they are independent of ranking, i.e. whether the order of returned documents the measures would give the same values. Since we want to verify improvement of a ranking we introduce modifications of precision and recall: P@n and R@n. P@n is precision at n positions, where n is the number of documents in a query result marked by user as pertinent. As expressed by Formula 2 P@n gives the fraction of pertinent documents within top n documents in a ranking.

$$P@n = \frac{|pertinent(top(n))|}{n} \quad . \tag{2}$$

P@n not always would give different values for two different rankings. In order to always be able to tell whether two rankings are different at first n position we introduce R@n. This measure is calculated according to Formula 3. It compares sums of ranks of top n pertinent documents of an ideal ranking (only pertinent documents) and the measured one.

$$R@n = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i} = \frac{n(n+1)}{2\sum_{i=1}^{n} r_i} \quad . \tag{3}$$

5.2 Results

Verification tests included evaluating pertinence of documents in 28 query results by 17 users (workers and students of Wrocaw University of Technology). Table 1 shows P@n and R@n values for ranking and reranking of those query results. Comparing P@n for standard *tf-idf* ranking (Solr) and DAC clustering reranking (DAC) it shows that the latter on average gives better results by circa 3%. The same improvement with reranking can be observed for R@n measure. In order

to prove statistical significance of the results statistical hypothesis testing has been performed. Since the used test: *t-test* is a parametric one, first *Kolmogorov-Smirnov test* for normality of the distribution had been carried out. Then the t-test p-values (<0,05) allowed to conclude that the null hypothesis is false for both P@n and R@n.

Table 1. Results from the experiment of comparing P@n precision and R@n recall of user query result orders: standard ranking based on term weighting (Solr) and improved ranking based on using semantic information and DAC graph clustering (DAC). P@n and R@n values are calculated on the basis of user evaluations of result document pertinence (relevance). T-test p-values are presented to prove statistical significance of the results.

query	P@n(Solr)	P@n(DAC)	R@n(Solr)	R@n(DAC)
skiing in austria	0	0	0,5	0,5
polish food	0	0	0,14	0,14
formula and one	0,32	0,37	0,24	0,24
e-learning	0,2	0,2	0,1	0,1
multimedia_video	1	1	1	1
hendrix jimi	0,2	0,2	0,36	0,36
tsunami	0,17	0,17	0,26	0,28
u2 music	0,8	0,8	0,47	0,5
semantic web	0,71	0,71	0,58	0,7
social AND network_social networking	0,78	0,67	0,79	0,7
semantic AND web	0,56	0,56	0,56	0,51
ajax	0,29	0,43	0,46	0,44
record AND video	0,13	0,13	0,26	0,27
skype	0,29	0,29	0,19	0,18
garden_flowers	0,64	0,64	0,59	0,59
pregnancy_baby	0,58	0,58	0,57	0,57
groove music	0,4	0,2	0,41	0,43
nobel	0,5	0,5	0,53	0,53
liverpool	0,65	0,73	0,68	0,71
god_novel	0	0	0,08	0,05
radiohead_radiohead	0,57	0,57	0,54	0,57
upgrade ubuntu	0,43	0,43	0,47	0,51
rock_music	0,47	0,6	0,5	0,66
CSS	0,25	0,25	0,18	0,2
validator	0	0,2	0,15	0,22
google labs	0,4	0,6	0,42	0,51
nintendo	0,57	0,71	0,45	0,7
kung fu	0,71	0,86	0,68	0,68
average	0,414	0,442	0,434	0,459
Kolmogorov-Smirnov test	1,66E-06	1,66E-06	2,83E-07	5,63E-07
t-test	0,048		0,027	

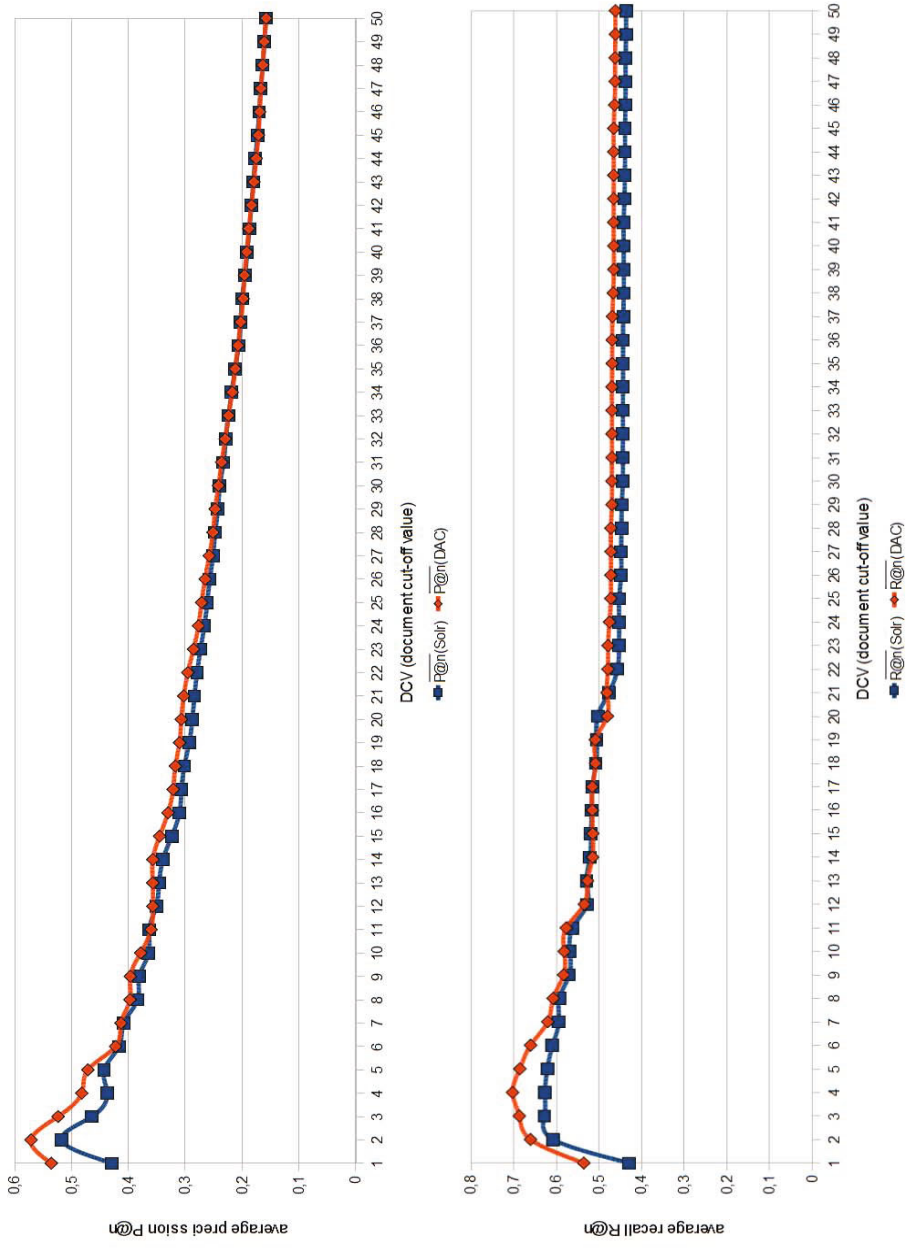


Fig. 3. Comparison of average P@n precision and average R@n recall of 28 user query result orders: standard ranking based on term weighting (Solr) and improved ranking based on using semantic information and DAC graph clustering (DAC). Each DCV correspond to number of top documents for which average P@n and average R@n are calculated.

5.3 DCV Diagrams

For additional visualization of P@n and R@n results for ranking and reranking DCV diagrams, inspired by [7], are used. In the cited paper the effectiveness of eight search engines was evaluated using precision and recall measures. But the measures were not calculated once for each query result but for every subset of top n documents, where n - the document cut-off value (DCV) - ranged from 1 to 20 (relevance evaluated by users for only 20 first documents). Since in our tests users evaluated all result documents and and average query result contained 58,8 documents, we use DCV range from 1 to 50.

As shown on Fig. 3 both P@n and R@n are significantly better for lower DCVs, which is most important for a user. For DCVs>30 P@n values for both rankings are equal, which possibility was discussed earlier as a reason for introducing R@n. And as supposed with DCV approaching infinity the difference between R@n values for both rankings is constant, but still different from zero.

6 Conclusions

Semanic search is not a new trend in Web IR, but still it is a "hot" topic. More and more services claim to use semantics, not only within search domain. But even within search domain using semantics may mean different things. One thing sure is that semantic search is an enhanced version of the traditional, Web 1.0, keyword based search. In this paper the semantic enhancement of a Web search concerns using Web objects relationships as and additional information for bringing more relevant results.

Because of the availability of the semantic information source, which is a blogging platform, relationships of three objects types are considered: document, author and concept. The object and the relationships are modeled with a DAC graph, which gives an augmentation to a traditional model - document graph. Example application of the DAC model is improvement of a *tf-idf* ranking. In order to obtain reranking - a better ranking - clustering of DAC graph is performed. Then new ranks for the clusters are calculated and they form a new ranking.

This new ranking depends heavily on relationships modeled by DAC. The little improvement of ranking in the experiment (3% better precision and recall) can be explained by the amount of the semantic information available. As pointed out in section 4, author relationships based on XFN links form rather sparse social graph. Another problem which may hugely impact clustering results and the reranking are concept relationships. Using Wordnet as a source of concept relationship values required mapping blog post tags to Wordnet synsets. In the experiment tags are treated as nouns and only its first meanings are taken into account. This limitation, imposed by performance issues, restricts the size of hypernym hiearchy and thus many concept relationships (paths in the tree) may not be found. This is the main aspect to be taken into account in further research.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American 284, 28–37 (2001)
2. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM 46, 604–632 (1998)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
4. Kopel, M., Zgrzywa, A.: The consistency and conformance of web document collection based on heterogeneous DAC graph. New Frontiers in Applied Artificial Intelligence, 321–330 (2008)
5. Kopel, M., Zgrzywa, A.: Application of Agent-Based personal web of trust to local document ranking. Agent and Multi-Agent Systems: Technologies and Applications, 288–297 (2007)
6. Dongen, S.: Graph clustering by flow simulation. PhD dissertation. University of Utrecht (2000)
7. Gordon, M., Pathak, P.: Finding information on the world wide web: the retrieval effectiveness of search engines. Inf. Process. Manage. 35, 141–180 (1999)

Data Filling Approach of Soft Sets under Incomplete Information

Hongwu Qin, Xiuqin Ma, Tutut Herawan, and Jasni Mohamad Zain

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
qhwump@gmail.com, xueener@yahoo.com.cn,
tutut@ump.edu.my, jasni@ump.edu.my

Abstract. Incomplete information in a soft set restricts the usage of the soft set. To make the incomplete soft set more useful, in this paper, we propose a data filling approach for incomplete soft set in which missing data is filled in terms of the association degree between the parameters when stronger association exists between the parameters or in terms of the probability of objects appearing in the mapping sets of parameters when no stronger association exists between the parameters. An illustrative example is employed to show the feasibility and validity of our approach in practical applications.

Keywords: Soft sets, Incomplete soft sets, Data filling, Association degree.

1 Introduction

In 1999, Molodtsov [1] proposed soft set theory as a new mathematical tool for dealing with vagueness and uncertainties. At present, work on the soft set theory is progressing rapidly and many important theoretical models have been presented, such as soft groups [2], soft rings [3], soft semirings [4], soft ordered semigroup [5] and exclusive disjunctive soft sets [6]. The research on fuzzy soft set has also received much attention since its introduction by Maji et al. [7]. Several extension models including intuitionistic fuzzy soft sets [8], interval-valued fuzzy soft sets [9] and interval-valued intuitionistic fuzzy soft set [10] are proposed in succession. At the same time, researchers have also successfully applied soft sets to deal with some practical problems, such as decision making [11-14], economy forecasting [15], maximal association rules mining [16], etc.

The soft sets mentioned above, either in theoretical study or practical applications are based on complete information. However, incomplete information widely exists in practical problems. For example, an applicant perhaps misses age when he/she fills out an application form. Missing or unclear data often appear in questionnaire due to the fact that attendees give up some questions or can not understand the meaning of questions well. In addition, other reasons like mistakes in the process of measuring and collecting data, restriction of data collecting also can cause unknown or missing data. Hence, soft sets under incomplete information become incomplete soft sets. In order to handle incomplete soft sets, new data processing methods are required.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 302–311, 2011.

Yan and Zhi [17] initiated the study on soft sets under incomplete information. They put forward improved data analysis approaches for standard soft sets and fuzzy soft sets under incomplete information, respectively. For crisp soft sets, the decision value of an object with incomplete information is calculated by weighted-average of all possible choice values of the object, and the weight of each possible choice value is decided by the distribution of other available objects. Incomplete data in fuzzy soft sets is predicted based on the method of average probability. However, there is inherent deficiency in their method. For crisp soft sets, directly calculating the decision value of an object with incomplete information makes the method only applicable to decision making problems. During the process of data analysis the soft sets keep invariable, in other words the missing data is still missing. Therefore, the soft sets can not be used in other fields but decision making.

Intuitively, there are two methods which can be used to overcome the deficiency in [17]. The simplest method is deletion that the objects with incomplete data will be deleted directly from incomplete soft sets. This method, however, probably makes valuable information missing. Another method is data filling, that is, the incomplete data will be estimated or predicted based on the known data. Data filling converts an incomplete soft set into a complete soft set, which makes the soft set more useful. So far, few researches focus on data filling approaches for incomplete soft sets.

In this paper, we propose a data filling approach for incomplete soft sets. We analyze the relations between the parameters and define the notion of association degree to measure the relations. In our method, we give priority to the relations between the parameters due to its higher reliability. When the mapping set of a parameter includes incomplete data, we firstly look for another parameter which has the stronger association with the parameter. If another parameter is found, the missing data in the mapping set of the parameter will be filled according to the value in the corresponding mapping set of another parameter. If no parameter has the stronger association with the parameter, the missing data will be filled in terms of the probability of objects appearing in the mapping set of the parameter. There are two main contributions in this work. First, we present the applicability of the data filling method to handle incomplete soft sets. Second, we introduce the relation between parameters to fill the missing data.

The rest of this paper is organized as follows. The following section presents the notions of soft sets and incomplete soft sets. Section 3 analyzes the relation between the parameters of soft set and defines the notion of association degree to measure the relation. In Section 4, we present our algorithm for filling the missing data and give an illustrative example. Finally, conclusions are given in Section 5.

2 Preliminaries

Let U be an initial universe of objects, E be the set of parameters in relation to objects in U, $P(U)$ denote the power set of U. The definition of soft set is given as follows.

Definition 2.1 ([1]). A pair (F, E) is called *a soft set* over U, where F is a mapping given by

$$F : E \to P(U)$$

From definition, a soft set (F, E) over the universe U is a parameterized family of subsets of the universe U, which gives an approximate description of the objects in U. For any parameter $e \in E$, the subset $F(e) \subseteq U$ may be considered as the set of e-approximate elements in the soft set (F, E).

Example 1. Let us consider a soft set (F, E) which describes the "attractiveness of houses" that Mr. X is considering to purchase. Suppose that there are six houses in the univers $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ under consideration and $E = \{e_1, e_2, e_3, e_4, e_5\}$ is the parameter set, where $e_i (i = 1,2,3,4,5)$ stands for the parameters "beautiful", "expensive", "cheap", "good location" and "wooden" respectively. Consider the mapping $F : E \to P(U)$ given by "houses (.)", where (.) is to be filled in by one of parameters $e \in E$. Suppose that $F(e_1) = \{h_1, h_3, h_6\}$, $F(e_2) = \{h_1, h_2, h_3, h_6\}$, $F(e_3) = \{h_4, h_5\}$, $F(e_4) = \{h_1, h_2, h_6\}$, $F(e_5) = \{h_5\}$. Therefore, $F(e_1)$ means "houses (beautiful)", whose value is the set $\{h_1, h_3, h_6\}$.

In order to facilitate storing and dealing with soft set, the binary tabular representation of soft set is often given in which the rows are labeled by the object names and columns are labeled by the parameter names, and the entries are $F(e_j)(x_i), (e_j \in E, x_i \in U, j = 1,2,..m, x = 1,2,..n)$. If $x_i \in F(e_j)$, then $F(e_j)(x_i) = 1$, otherwise $F(e_j)(x_i) = 0$. Table 1 is the tabular representation of the soft set (F, E) in Example 1.

Table 1. Tabular representation of the soft set (F, E)

U	e_1	e_2	e_3	e_4	e_5
h_1	1	1	0	1	0
h_2	0	1	0	1	0
h_3	1	1	0	0	0
h_4	0	0	1	0	0
h_5	0	0	1	0	1
h_6	1	1	0	1	0

Definition 2.2. A pair (F, E) is called *an incomplete soft set* over U, if there exists $x_i \in U (i = 1,2...,n)$ and $e_j \in E (j = 1,2...,m)$, making $x_i \in F(e_j)$ unknown, that is, $F(e_j)(x_i) = null$.

In tabular representation, null is represented by "*".

Example 2. Assume a community college is recruiting some new teachers and there are 8 persons applied for the job. Let us consider a soft set (F, E) which describes the "capability of the candidates". The universe $U = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$ and $E = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ is the parameter set, where $e_i (i = 1,2,3,4,5,6)$ stands for the parameters "experienced", "young age", "married", "the highest academic degree is Doctor", "the highest academic degree is Master" and "studied abroad" respectively. Consider the mapping $F : E \to P(U)$ given by "candidates (.)", where (.) is to be filled in by one of parameters $e \in E$. Suppose that

$$F(e_1) = \{c_1, c_2, c_5, c_7\},\ F(e_2) = \{c_3, c_4, c_6\},\ F(e_3) = \{c_1, c_5, c_7, c_8\},$$
$$F(e_4) = \{c_2, c_4, c_5, c_8\},\quad F(e_5) = \{c_1, c_3, c_6, c_7\},\quad F(e_6) = \{c_8\}.$$

Therefore, $F(e_1)$ means "candidates (experienced)", whose value is the set $\{c_1, c_2, c_5, c_7\}$. Unfortunately, several applicants missed some information. As a result, the soft set (F, E) becomes an incomplete soft set. Table 2 is the tabular representation of the incomplete soft set (F, E). If $c_j \in F(e_i)$ is unknown, $F(e_i)(c_j) = '*'$, where $F(e_i)(c_j)$ are the entries in Table 2.

Table 2. Tabular representation of the incomplete soft set (F, E)

U	e_1	e_2	e_3	e_4	e_5	e_6
c_1	1	0	1	0	1	0
c_2	1	0	0	1	0	0
c_3	0	1	0	0	1	0
c_4	0	1	*	1	0	*
c_5	1	0	1	1	0	0
c_6	0	1	0	0	*	0
c_7	1	*	1	0	1	0
c_8	0	0	1	1	0	0

3 Association Degree between Parameters in an Incomplete Soft Set

So far, few research focus on the associations between parameters in the soft sets. Actually, for one object, there always exist some obvious or hidden associations between parameters. This is just like for a person, as we know, the attribute weight has some certain relation with the attribute height.

Let us reconsider Example 1 and Example 2. There are many obvious associations in the two examples. In Example 1, it is easy to find that if a house is expensive, the house is not cheap, vice versa. There is inconsistent association between parameter "expensive" and parameter "cheap". Generally speaking, if a house is beautiful or has a good location, the house is expensive. There is consistent association between parameter "beautiful" and parameter "expensive" or between parameter "good location" and parameter "expensive". Similarly, in Example 2, there is obvious inconsistent association between parameter "the highest academic degree is Doctor" and parameter "the highest academic degree is Master". A candidate has only one highest academic degree. We can also find that if a candidate is experienced or has been married, in general, he/she is not young. There is inconsistent association between parameter "experienced" and parameter "young age" or between parameter "married" and parameter "young age".

These associations reveal the interior relations of an object. In a soft set, these associations between parameters will be very useful for filling incomplete data. If we have already found that parameter e_i is associated with parameter e_j and there are missing data in $F(e_i)$, we can filling the missing data according to the corresponding data in $F(e_j)$ based on the association between e_i and e_j. To measure these associations, we define the notion of association degree and some relative notions.

Let U be a universe set and E be a set of parameters. U_{ij} denotes the set of objects that have specified values 0 or 1 both on parameter e_i and parameter e_j such that

$$U_{ij} = \{x \mid F(e_i)(x) \neq '*' \quad and \quad F(e_j)(x) \neq '*', x \in U\}$$

In other words, U_{ij} stands for the set of objects that have known data both on e_i and e_j. Based on U_{ij}, we have the following definitions.

Definition 3.1. Let E be a set of parameters and $e_i, e_j \in E, (i, j = 1,2,...m)$. *Consistent Association Number* between parameter e_i and parameter e_j is denoted by CN_{ij} and defined as

$$CN_{ij} = \left|\{x \mid F(e_i)(x) = F(e_j)(x), x \in U_{ij}\}\right|$$

where m denotes the number of parameters, $|.|$ denotes the cardinality of set.

Definition 3.2. Let E be a set of parameters and $e_i, e_j \in E, (i, j = 1,2,...m)$. *Consistent Association Degree* between parameter e_i and parameter e_j is denoted by CD_{ij} and defined as

$$CD_{ij} = \frac{CN_{ij}}{|U_{ij}|}$$

Obviously, the value of CD_{ij} is in [0, 1]. Consistent Association Degree measures the extent to which the value of parameter e_i keeps consistent with that of parameter e_j over U_{ij}.

Similarly, we can define *Inconsistent Association Number* and *Inconsistent Association Degree* as follows.

Definition 3.3. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Inconsistent Association Number* between parameter e_i and parameter e_j is denoted by IN_{ij} and defined as

$$IN_{ij} = \left|\left\{x \mid F(e_i)(x) \neq F(e_j)(x), x \in U_{ij}\right\}\right|$$

Definition 3.4. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Inconsistent Association Degree* between parameter e_i and parameter e_j is denoted by ID_{ij} and defined as

$$ID_{ij} = \frac{IN_{ij}}{\left|U_{ij}\right|}$$

Obviously, the value of ID_{ij} is also in [0, 1]. Inconsistent Association Degree measures the extent to which parameters e_i and e_j is inconsistent.

Definition 3.5. Let E be a set of parameters and $e_i, e_j \in E$, $(i, j = 1,2,...m)$. *Association Degree* between parameter e_i and parameter e_j is denoted by D_{ij} and defined as

$$D_{ij} = \max\left\{CD_{ij}, ID_{ij}\right\}$$

If $CD_{ij} > ID_{ij}$, then $D_{ij} = CD_{ij}$, it means most of objects over U_{ij} have consistent values on parameters e_i and e_j. If $CD_{ij} < ID_{ij}$, then $D_{ij} = ID_{ij}$, it means most of objects over U_{ij} have inconsistent values on parameters e_i and e_j. If $CD_{ij} = ID_{ij}$, it means that there is the lowest association degree between parameters e_i and e_j.

Property 3.1. For any parameters and e_j, $D_{ij} \geq 0.5$. $(i, j = 1,2,...m)$.

Proof. For any parameters e_i and e_j, from the definitions of CD_{ij} and ID_{ij}, we have

$$CD_{ij} + ID_{ij} = 1.$$

Therefore, at least one of CD_{ij} and ID_{ij} is more than 0.5, namely, $D_{ij} = \max\left\{CD_{ij}, ID_{ij}\right\} \geq 0.5$. □

Definition 3.6. Let E be a set of parameters and $e_i \in E$ $(i = 1,2,...m)$. *Maximal Association Degree* of parameter e_i is denoted by D_i and defined as

$$D_i = \max D_{ij}, \quad \mathrm{j} = 1,2,...\mathrm{m}.$$

where m is the number of parameters.

4 The Algorithm for Data Filling

In terms of the analysis in the above section, we can propose the data filling method based on the association degree between the parameters. Suppose the mapping set $F(e_i)$ of parameter e_i includes missing data. At first, calculate association degrees between parameter e_i and each of other parameters respectively over existing complete information, and then find the parameter e_j which has the maximal association degree with parameter e_i. Finally the missing data in $F(e_i)$ will be filled according to the corresponding data in mapping set $F(e_j)$. However, sometimes a parameter perhaps has a lower maximal association degree, that is, the parameter has weaker association with other parameters. In this case, the association is not reliable any more and we have to find other methods. Inspired by the data analysis approach in [17], we can use the probability of objects appearing in the $F(e_i)$ to fill the missing data. In our method we give priority to the association between the parameters instead of the probability of objects appearing in the $F(e_i)$ to fill the missing data due to the fact that the relation between the parameters are more reliable than that between the objects in soft set. Therefore, we can set a threshold, if the maximal association degree equals or exceeds the predefined threshold, the missing data in $F(e_i)$ will be filled according to the corresponding data in $F(e_j)$, or else the missing data will be filled in terms of the probability of objects appearing in the $F(e_i)$. Fig. 1 shows the details of the algorithm.

In order to make the computation of association degree easier, we construct an association degree table in which rows are labeled by the parameters including missing data and columns are labeled by all of the parameters in parameter set, and the entries are association degree D_{ij}. To distinguish the inconsistent association degree from consistent degree, we add a minus sign before the inconsistent association degree.

Example 3. Reconsider the incomplete soft set (F, E) in Example 2. There are missing data in $F(e_2)$, $F(e_3)$, $F(e_5)$ and $F(e_6)$. We will fill the missing data in (F, E) by using Algorithm 1. Firstly, we construct an association degree table as Table 3.

For parameter e_2, we can see from the table, the association degree $D_{21} = 0.86$, $D_{23} = 0.83$, $D_{24} = 0.71$, $D_{25} = 0.67$, $D_{26} = 0.67$, where D_{21}, D_{23} and D_{24} are

Algorithm 1.

1. Input the incomplete soft set (F, E).
2. Find e_i, which includes missing data $F(e_i)(x)$.
3. Compute D_{ij}, $j = 1,2,...,m$, where m is the number of parameters in E.
4. Compute maximal association degree D_i.
5. If $D_i \geq \lambda$, find the parameter e_j which has the maximal association degree D_i with parameter e_i.
6. If there is consistent association between e_i and e_j, $F(e_i)(x) = F(e_j)(x)$. If there is inconsistent association between e_i and e_j, $F(e_i)(x) = 1 - F(e_j)(x)$.
7. If $D_i < \lambda$, compute the probabilities P_1 and P_0 that stand for object x belongs to and does not belong to $F(e_i)$, respectively.

$$P_1 = \frac{n_1}{n_1 + n_0}, \quad P_0 = \frac{n_0}{n_1 + n_0}$$

where n_1 and n_0 stand for the number of objects that belong to and does not belong to $F(e_i)$, respectively.

8. If $P_1 > P_0$, $F(e_i)(x) = 1$. If $P_0 > P_1$, $F(e_i)(x) = 0$. If $P_1 = P_0$, 0 or 1 may be assigned to $F(e_i)(x)$.
9. If all of the missing data is filled, algorithm end, or else go to step 2.

Fig. 1. The algorithm for data filling

from inconsistent association degree, D_{25} and D_{26} are from consistent association degree. The maximal association degree $D_2 = 0.86$. We set the threshold $\lambda = 0.8$. Therefore, in terms of the Algorithm 1, we can fill $F(e_2)(c_7)$ according to $F(e_1)(c_7)$. Because $F(e_1)(c_7) = 1$ and there is inconsistent association between parameters e_2 and e_1, so we fill 0 into $F(e_2)(c_7)$. Similarly, we can fill 0, 1 into $F(e_3)(c_4)$ and $F(e_5)(c_6)$ respectively.

Table 3. Association degree table for incomplete soft set (F, E)

	e_1	e_2	e_3	e_4	e_5	e_6
e_2	-0.86	-	-0.83	-0.71	0.67	0.67
e_3	0.71	-0.83	-	0.57	0.5	-0.57
e_5	0.57	0.67	0.5	-1	-	0.5
e_6	-0.57	0.67	0.57	0.57	0.5	-

For parameter e_6, we have the maximal association degree $D_6 = 0.67 < \lambda$. That means there is not reliable association between parameter e_6 and other parameters. So we can not fill the data $F(e_6)(c_4)$ according to other parameters. In terms of the steps 8 and 9 in Algorithm 1, we have $P_0 = 1$, $P_1 = 0$. Therefore, we fill 0 into $F(e_6)(c_4)$. Table 4 shows the tabular representation of the filled soft set (F, E).

Table 4. Tabular representation of the incomplete soft set (F, E)

U	e_1	e_2	e_3	e_4	e_5	e_6
c_1	1	0	1	0	1	0
c_2	1	0	0	1	0	0
c_3	0	1	0	0	1	0
c_4	0	1	0	1	0	0
c_5	1	0	1	1	0	0
c_6	0	1	0	0	1	0
c_7	1	0	1	0	1	0
c_8	0	0	1	1	0	0

5 Conclusion

In this paper, we propose a data filling approach for incomplete soft sets. We analyze the relations between the parameters and define the notion of association degree to measure the relations. If the mapping set of a parameter includes incomplete data, we firstly look for another parameter which has the stronger association with the parameter. If another parameter is found, the missing data in the mapping set of the parameter will be filled according to the value in the corresponding mapping set of another parameter. If no parameter has the stronger association with the parameter, the missing data will be filled in terms of the probability of objects appearing in the mapping set of the parameter. We validate the method by an example and draw conclusion that data filling method is applicable to handle incomplete soft sets and the relations between parameters can be applied to fill the missing data. The method can be used to handle various applications involved incomplete soft sets.

Acknowledgments. This work was supported by PRGS under the Grant No. GRS100323, Universiti Malaysia Pahang, Malaysia.

References

1. Molodtsov, D.: Soft set theory_First results. Computers and Mathematics with Applications 37, 19–31 (1999)
2. Aktas, H., Cagman, N.: Soft sets and soft groups. Information Sciences 177, 2726–2735 (2007)

3. Acar, U., Koyuncu, F., Tanay, B.: Soft sets and soft rings. Computers and Mathematics with Applications 59, 3458–3463 (2010)
4. Feng, F., Jun, Y.B., Zhao, X.: Soft semirings. Computers and Mathematics with Applications 56, 2621–2628 (2008)
5. Jun, Y.B., Lee, K.J., Khan, A.: Soft ordered semigroups. Math. Logic Quart. 56, 42–50 (2010)
6. Xiao, Z., Gong, K., Xia, S., Zou, Y.: Exclusive disjunctive soft sets. Computers and Mathematics with Applications 59, 2128–2137 (2010)
7. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9, 589–602 (2001)
8. Maji, P.K., Biswas, R., Roy, A.R.: Intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 9, 677–692 (2001)
9. Yang, X.B., Lin, T.Y., Yang, J., Dongjun, Y.L.A.: Combination of interval-valued fuzzy set and soft set. Computers and Mathematics with Applications 58, 521–527 (2009)
10. Jiang, Y., Tang, Y., Chen, Q., Liu, H., Tang, J.: Interval-valued intuitionistic fuzzy soft sets and their properties. Computers and Mathematics with Applications 60, 906–918 (2010)
11. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
12. Feng, F., Jun, Y.B., Liu, X., Li, L.: An adjustable approach to fuzzy soft set based decision making. Journal of Computational and Applied Mathematics 234, 10–20 (2010)
13. Feng, F., Li, Y., Leoreanu-Fotea, V.: Application of level soft sets in decision making based on interval-valued fuzzy soft sets. Computers and Mathematics with Applications 60, 1756–1767 (2010)
14. Jiang, Y., Tang, Y., Chen, Q.: An adjustable approach to intuitionistic fuzzy soft sets based decision making. Applied Mathematical Modelling 35, 824–836 (2011)
15. Xiao, Z., Gong, K., Zou, Y.: A combined forecasting approach based on fuzzy soft sets. Journal of Computational and Applied Mathematics 228, 326–333 (2009)
16. Herawan, T., Mat Deris, M.: A soft set approach for association rules mining. Knowledge-Based Systems 24, 186–195 (2011)
17. Zou, Y., Xiao, Z.: Data analysis approaches of soft sets under incomplete information. Knowledge-Based Systems 21, 941–945 (2008)

Empirical Comparison of Bagging Ensembles Created Using Weak Learners for a Regression Problem

Karol Bańczyk[1], Olgierd Kempa[2], Tadeusz Lasota[2], and Bogdan Trawiński[1]

[1] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[2] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wrocław, Poland
{karol.banczyk,tadeusz.lasota}@wp.pl, olgierd_kempa@vp.pl,
bogdan.trawinski@pwr.wroc.pl

Abstract. The experiments, aimed to compare the performance of bagging ensembles using three different test sets composed of base, out-of-bag, and 30% holdout instances were conducted. Six weak learners including conjunctive rules, decision stump, decision table, pruned model trees, rule model trees, and multilayer perceptron, implemented in the data mining system WEKA, were applied. All algorithms were employed to real-world datasets derived from the cadastral system and the registry of real estate transactions, and cleansed by property valuation experts. The analysis of the results was performed using recently proposed statistical methodology including nonparametric tests followed by post-hoc procedures designed especially for multiple n×n comparisons. The results showed the lowest prediction error with base test set only in the case of model trees and a neural network.

Keywords: ensemble models, bagging, out-of-bag, property valuation, WEKA.

1 Introduction

Bagging ensembles, which besides boosting belong to the most popular multi-model techniques have been focused attention of many researchers for last fifteen years. Bagging, which stands for bootstrap aggregating, devised by Breiman [2] is one of the most intuitive and simplest ensemble algorithms providing a good performance. Diversity of learners is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the original training set. So obtained training data subsets, called also bags, are used then to train different classification and regression models. Finally, individual learners are combined through an algebraic expression, such as minimum, maximum, sum, mean, product, median, etc. [19]. Theoretical analyses and experimental results proved benefits of bagging especially in terms of stability improvement and variance reduction of learners for both classification and regression problems [3], [8], [9].

This collection of methods combines the output of the machine learning systems, in literature called "weak learners" in due to its performance [20], from the group of learners in order to get smaller prediction errors (in regression) or lower error rates (in

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 312–322, 2011.

classification). The individual estimator must provide different patterns of generalization, thus the diversity plays a crucial role in the training process. Otherwise, the ensemble, called also committee, would end up having the same predictor and provide as good accuracy as the single one. It was proved that the ensemble performs better when each individual machine learning system is accurate and makes error on the different instances at the same time.

The size of bootstrapped replicas in bagging usually is equal to the number of instances in an original dataset and the base dataset (Base) is commonly used as a test set for each generated component model. However, it is claimed it leads to an optimistic overestimation of the prediction error. So, as test error out-of-bag samples (OoB) are applied, i.e. those included in the Base dataset but not drawn to respective bags. These, in turn may cause a pessimistic underestimation of the prediction error. In consequence, correction estimators are proposed which are linear combinations of errors provided by Base and OoB test sets [4], [7].

So far we have investigated several methods to construct regression models to assist with real estate appraisal: evolutionary fuzzy systems, neural networks, decision trees, and statistical algorithms using MATLAB, KEEL, RapidMiner, and WEKA data mining systems [13], [15], [17]. We studied also ensemble models created applying various fuzzy systems, neural networks, support vector machines, regression trees, and statistical regression [14], [16], [18].

The main goal of the study presented in this paper was to investigate the usefulness of 17 machine learning algorithms available as Java classes in WEKA software to create ensemble models providing better performance than their single base models. The algorithms were applied to real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transaction obtained from a cadastral system. In the paper a part of results is presented comprising six of eight models which revealed prediction error reduction of bagging ensembles compared to base models. The models were built using weak learners including three simple ones: conjunctive rules, decision stump, and decision table as well as pruned model trees, rule model trees, and multilayer perceptron. The experiments aimed also to compare the impact of three different test sets composed of base, out-of-bag, and 30% holdout instances on the performance of bagging ensembles.

2 Algorithms Used and Plan of Experiments

The investigation was conducted with an experimental multi-agent system implemented in Java using learner classes available in WEKA library [44]. WEKA *(Waikato Environment for Knowledge Analysis),* a non-commercial and open-source data mining system, comprises tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [5], [21]. WEKA encompasses many algorithms for classification and numeric prediction, i.e. regression problems. The latter is interpreted as prediction of a continuous class.

In our experiments we employed 17 learners taken from WEKA library to create bagging ensembles to examine how they improve the performance of models to assist with real estate appraisal compared to single base models. However, following nine

ensemble models did not provide lower prediction error: *GaussianProcesses, IBk, IsotonicRegression, KStar, LeastMedSq, LinearRegression, PaceRegression, REPTree, SMOreg*. In the case of *ConjunctiveRule, DecisionStump, DecisionTable, M5P, M5Rules, MultilayerPerceptron, LWL, RBFNetwork* learners bagging ensembles revealed better performance. Due to the limited space the results referring to only six first ones are presented in the paper, all belong to weak learners.

CJR – ConjunctiveRule. This class implements a single conjunctive rule learner. A rule consists of antecedents combined with the operator AND and the consequent for the classification/regression. If the test instance is not covered by this rule, then it is predicted using the default value of the data not covered by the rule in the training data. This learner selects an antecedent by computing the information gain of each antecedent and prunes the generated rule. For regression, the information is the weighted average of the mean-squared errors of both the data covered and not covered by the rule.

DST – DecisionStump. Class for building and using a decision stump. It builds one-level binary decision trees for datasets with a categorical or numeric class, dealing with missing values by treating them as a separate value and extending a third branch from the stump. Regression is done based on mean-squared error.

DTB – DecisionTable. Class for building and using a simple decision table majority classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation. An option uses the nearest-neighbor method to determine the class for each instance that is not covered by a decision table entry, instead of the table's global majority, based on the same set of features.

M5P – Pruned Model Tree. Implements routines for generating M5 model trees The algorithm is based on decision trees, however, instead of having values at tree's nodes, it contains a multivariate linear regression model at each node. The input space is divided into cells using training data and their outcomes, then a regression model is built in each cell as a leaf of the tree.

M5R – M5Rules. Generates a decision list for regression problems using separate-and-conquer. In each iteration it builds an model tree using M5 and makes the "best" leaf into a rule. The algorithm divides the parameter space into areas (subspaces) and builds in each of them a linear regression model. It is based on M5 algorithm. In each iteration a M5 Tree is generated and its best rule is extracted according to a given heuristic. The algorithm terminates when all the examples are covered.

MLP – MultiLayerPerceptron. A Classifier that uses backpropagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units).

The dataset used in experiments was drawn out from a rough dataset containing above 50 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within eleven years from 1998 to 2008. In this period most transactions were made with non-market prices when the council was selling flats to their current tenants on preferential terms. First of all, transactional records referring to residential premises sold at market prices were selected. Then the dataset was confined to sales transaction data of apartments built before 1997 and where the land was leased on terms of perpetual usufruct.

Five following features were pointed out as main drivers of premises prices: usable area of premises, age of a building, number of rooms in a flat, number of storeys in a building, and distance from the city centre. Hence, the final dataset counted 5303 records. Due to the fact that the prices of premises change substantially in the course of time, the whole 11-year dataset cannot be used to create data-driven models using machine learning . Therefore it was split into subsets covering individual years, and we might assume that within one year the prices of premises with similar attributes were roughly comparable. The sizes of one-year data subsets are given in Table 1.

Table 1. Number of instances in one-year datasets

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
269	477	329	463	530	653	546	580	677	575	204

Three series of experiments were conducted each for different arrangements of training and test sets for each one-year dataset separately. The base dataset comprised the whole one-year dataset in first two cases whereas it was composed of the greater part obtained in the result of the 70%/30% random split of a one-year dataset. On the basis of each base dataset 50 bootstrap replicates (bags) were created. These replicates were then used as training sets to generate models employing individual learners. In order to assess the predictive capability of each model three different test sets were used, namely the base dataset, out-of-bag, and 30% split denoted in the rest of the paper as Base, OoB, and 30%H respectively. Diagrams illustrating the respective experiments are shown in Fig. 1, 2, and 3. Normalization of data was performed using the min-max approach. As performance functions the root mean square error (RMSE) and the Correlation between predicted and actual values were used. As aggregation functions averages were employed. Preliminary tuning tests were accomplished using the trial and error method in order to determine the best parameter settings of each learner for each arrangement. In order to determine the performance of base single models 10-fold cross-validation experiments were conducted.

Statistical analysis of the results of experiments was performed using Wilcoxon signed rank tests and recently proposed procedures adequate for multiple comparisons of many learning algorithms over multiple datasets [6], [10], [11], [12]. Their authors argue that the commonly used paired tests i.e. parametric t-test and its nonparametric alternative Wilcoxon signed rank tests are not adequate when conducting multiple comparisons due to the so called multiplicity effect. They recommend following methodology. First of all the Friedman test or its more powerful derivative the Iman and Davenport test are carried out. Both tests can only inform the researcher about the presence of differences among all samples of results compared. After the null-hypotheses have been rejected he can proceed with the post-hoc procedures in order to find the particular pairs of algorithms which produce differences. They comprise Bonferroni-Dunn's, Holm's, and Hochberg's procedures in the case of 1×n comparisons and Nemenyi's, Shaffer's, and Bergmann-Hommel's procedures in the case of n×n comparisons. We used JAVA programs available on the web page of Research Group "Soft Computing and Intelligent Information Systems" at the University of Granada (http://sci2s.ugr.es/sicidm).

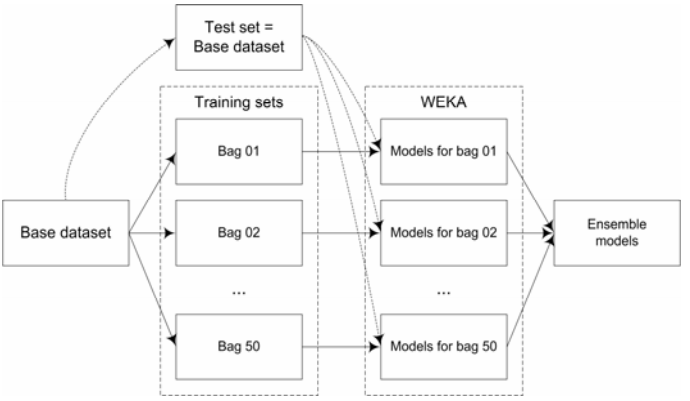


Fig. 1. Schema of the experiments with Base test set (Base)

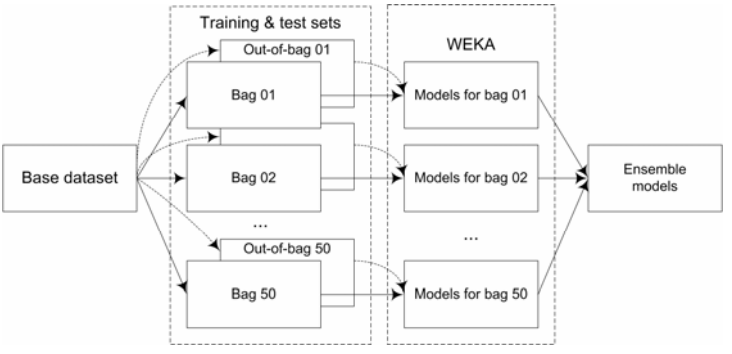


Fig. 2. Schema of the experiments with Out-of-bag test set (OoB)

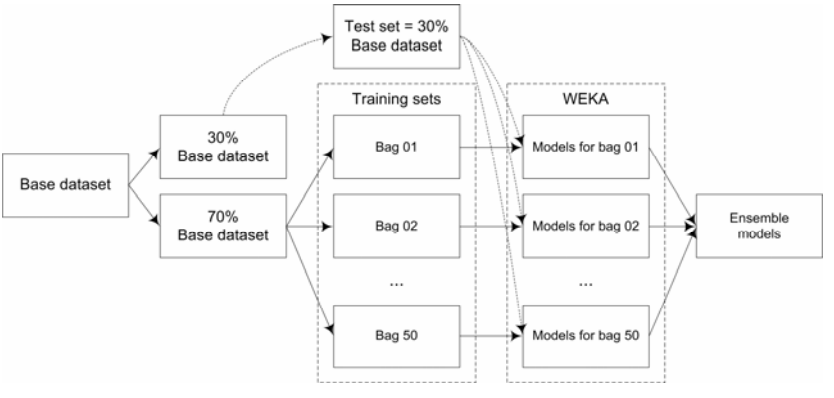


Fig. 3. Schema of the experiments with 30% holdout test set (30%H)

3 Results of Experiments

The performance of the ensemble models built by CJR, DST, DTB, M5P, M5R, and MLP in terms of RMSE and Correlation was presented in Figures 4-9 respectively. Each bar chart illustrates the relationship among the outcome of the models with Base, OoB, and 30%H test sets for successive one-year datasets. Only M5P, M5R, and MLP models confirm the observations claimed by many authors that Base test set provides optimistic and OoB pessimistic estimation of model accuracy, in turn 30%H one gives higher values of RMSE because training sets contained smaller numbers of instances. In the case of DST and DTB the models revealed the same performance for both Base and OoB test sets. Using Correlation between predicted and actual values exactly inverse relationship could be observed; for this performance measure the higher value the better.

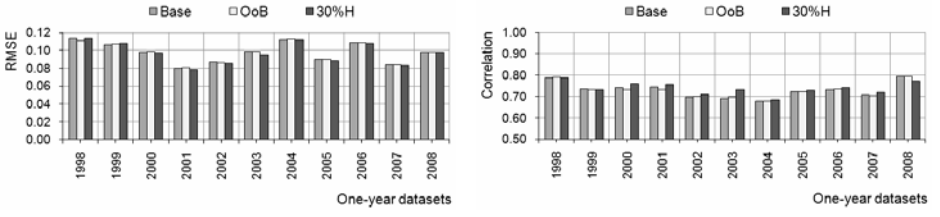


Fig. 4. Comparison of *ConjuctiveRule* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

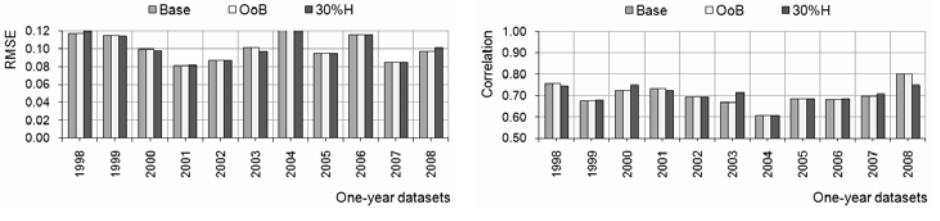


Fig. 5. Comparison of *DecisionStump* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

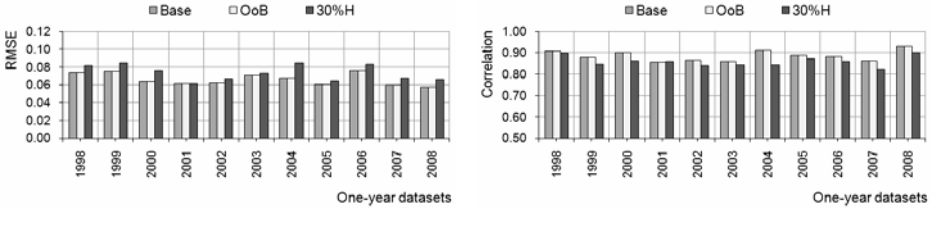


Fig. 6. Comparison of *DecisionTable* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

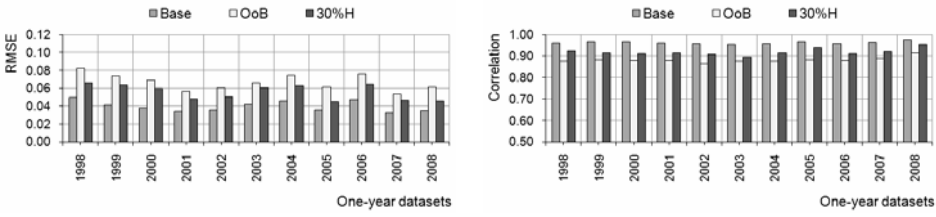


Fig. 7. Comparison of *M5P* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

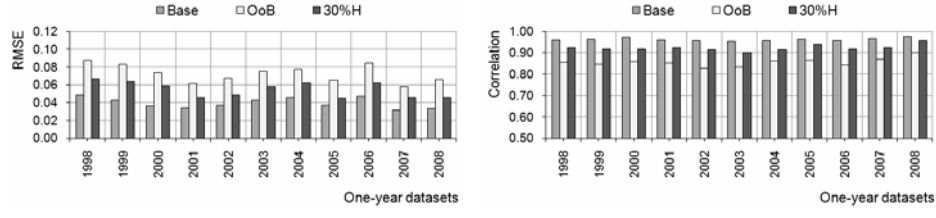


Fig. 8. Comparison of *M5Rules* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

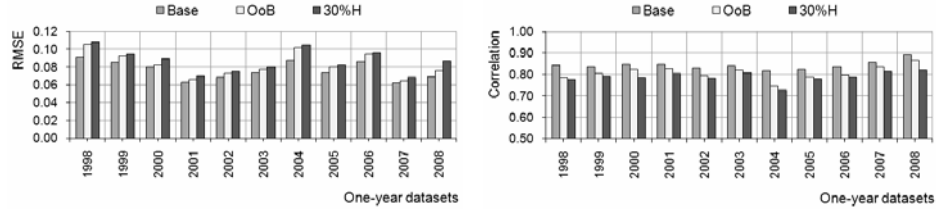


Fig. 9. Comparison of *MultiLayerPerceptron* bagging ensembles using different test sets, in terms of RMSE (left chart) and Correlation (right chart)

Table 2. Results of Wilcoxon tests for bagging ensembles with consecutive pairs of test sets

	Test sets	CJR	DST	DTB	M5P	M5R	MLP
RMSE	Base vs OoB	≈	≈	≈	+	+	+
	Base vs 30%H	–	≈	+	+	+	+
	OoB vs 30%H	≈	≈	+	–	–	+

In Table 2 the results of nonparametric Wilcoxon signed-rank test to evaluate the outcome of ensemble models using Base, OoB, and 30% test sets are presented. The zero hypothesis stated there were not significant differences in accuracy, in terms of RMSE, between given pairs of models. In Table 2 +, –, and ≈ denote that the first algorithm in a pair performed significantly better than, significantly worse than, or statistically equivalent to the second algorithm, respectively. Main outcome is as follows: for M5P, M5R, and MLP ensembles with Base test sets showed significantly better performance than those with OoB and 30%H test sets. For simple learners the results were not so clear, only for DTB the models with Base and OoB revealed

significantly lower RMSE than with 30%H, and CJR ensemble with 30%H was significantly better than with Base.

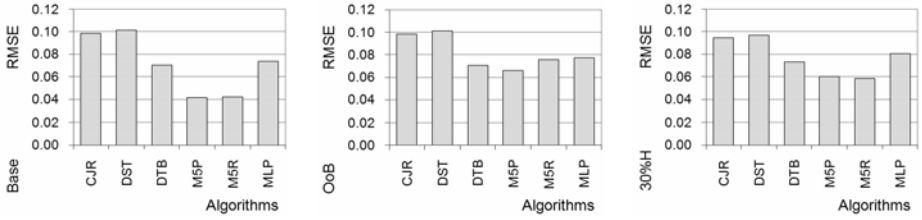


Fig. 10. Performance of bagging ensembles with Base test set (left chart), OoB test set (middle chart), and 30%H test set (right chart) for 2003 dataset

The experiments allowed also for the comparison of the ensembles build with individual algorithms. For illustration, in Figure 10 the results for models with Base, OoB, and 30% test sets for 2003 one-year dataset are shown, the charts for other datasets are similar. Statistical tests adequate to multiple comparisons were made for 6 algorithms altogether over all 11 one-year datasets. This tests, described in pervious section, were accomplished for models with Base, OoB, and 30%H test sets separately. The Friedman and Iman-Davenport tests were performed in respect of average ranks, which use χ^2 and F statistics. The calculated values of χ^2 statistics were equal to 52.92, 52.14, 54.06 for models with Base, OoB, and 30%H test sets respectively, and F statistics were 254.68, 182.49, 578.19, whereas the critical values at α=0.05 are $\chi^2(5)$=12.83 and F(5,50)=2.40. This means that there are significant differences between some models. Average ranks of individual ensembles are shown in Table 3, where the lower rank value the better model. Thus, we were justified in proceeding to post-hoc procedures. In Tables 4-6 adjusted p-values for Nemenyi, Holm, and Shaffer tests for n×n comparisons in terms of RMSE of bagging ensembles with Base, OoB, and 30%H test sets respectively are shown. In all tables the p-values less than α=0.05, indicating that respective models differ significantly, were marked with an italic font.

Table 3. Average rank positions of model performance in terms of RMSE

Test set	1st	2nd	3rd	4th	5th	6th
Base	M5R (1.45)	M5P (1.55)	DTB (3.00)	MLP (4.00)	CJR (5.09)	DST (5.91)
OoB	M5P (1.45)	DTB (1.64)	M5R (2.91)	MLP (4.00)	CJR (5.09)	DST (5.91)
30%H	M5R (1.18)	M5P (1.82)	DTB (3.00)	MLP (4.00)	CJR (5.00)	DST (6.00)

Table 4. Adjusted p-values for n×n comparisons in terms of RMSE of bagging ensembles with Base test sets showing 8 hypotheses rejected out of 15

Alg vs Alg	pUnadj	pNeme	pHolm	pShaf	pBerg
DST vs M5R	*2.35E-08*	*3.52E-07*	*3.52E-07*	*3.52E-07*	*3.52E-07*
DST vs M5P	*4.50E-08*	*6.75E-07*	*6.30E-07*	*4.50E-07*	*4.50E-07*
CJR vs M5R	*5.15E-06*	*7.73E-05*	*6.70E-05*	*5.15E-05*	*5.15E-05*
CJR vs M5P	*8.81E-06*	*1.32E-04*	*1.06E-04*	*8.81E-05*	*5.29E-05*
DST vs DTB	*2.66E-04*	*0.003984*	*0.002921*	*0.002656*	*0.001859*
M5R vs MLP	*0.001418*	*0.021275*	*0.014183*	*0.014183*	*0.009928*
M5P vs MLP	*0.002091*	*0.031371*	*0.018823*	*0.014640*	*0.009928*
CJR vs DTB	0.008765	0.131472	0.070119	0.061354	*0.035059*

Table 5. Adjusted p-values for n×n comparisons in terms of RMSE of bagging ensembles with OoB test sets showing 8 hypotheses rejected out of 15

Alg vs Alg	pUnadj	pNeme	pHolm	pShaf	pBerg
DST vs M5P	*2.35E-08*	*3.52E-07*	*3.52E-07*	*3.52E-07*	*3.52E-07*
DST vs DTB	*8.50E-08*	*1.28E-06*	*1.19E-06*	*8.50E-07*	*8.50E-07*
CJR vs M5P	*5.15E-06*	*7.73E-05*	*6.70E-05*	*5.15E-05*	*5.15E-05*
CJR vs DTB	*1.49E-05*	*2.23E-04*	*1.79E-04*	*1.49E-04*	*8.93E-05*
DST vs M5R	*1.69E-04*	*0.002542*	*0.001864*	*0.001694*	*0.001186*
M5P vs MLP	*0.001418*	*0.021275*	*0.014183*	*0.014183*	*0.009928*
DTB vs MLP	*0.003047*	*0.045702*	*0.027421*	*0.021328*	*0.012187*
CJR vs M5R	*0.006237*	*0.093555*	*0.049896*	*0.043659*	*0.024948*

Table 6. Adjusted p-values for n×n comparisons in terms of RMSE of bagging ensembles with 30%H test sets showing 7 hypotheses rejected out of 15

Alg vs Alg	pUnadj	pNeme	pHolm	pShaf	pBerg
DST vs M5R	*1.54E-09*	*2.31E-08*	*2.31E-08*	*2.31E-08*	*2.31E-08*
DST vs M5P	*1.59E-07*	*2.38E-06*	*2.22E-06*	*1.59E-06*	*1.59E-06*
CJR vs M5R	*1.70E-06*	*2.55E-05*	*2.21E-05*	*1.70E-05*	*1.70E-05*
CJR vs M5P	*6.65E-05*	*9.97E-04*	*7.98E-04*	*6.65E-04*	*3.99E-04*
DST vs DTB	*1.69E-04*	*0.002542*	*0.001864*	*0.001694*	*0.001186*
M5R vs MLP	*4.11E-04*	*0.006168*	*0.004112*	*0.004112*	*0.002879*
M5P vs MLP	*0.006237*	0.093555	0.056133	*0.043659*	*0.024948*
CJR vs DTB	*0.012172*	0.182573	0.097372	*0.085201*	*0.048686*

Following main observations could be done: M5P, M5R, and DTB revealed significantly better performance than CJR and DST models for all three test sets. There were not significant differences among M5P, M5R, and DTB and CJR, DST, and MLP models. M5P and M5R were significantly better than MLP for two test sets and DTB for one test set.

4 Conclusions and Future Work

The experiments, aimed to compare the performance of bagging ensembles using three different test sets composed of base, out-of-bag, and 30% holdout instances were conducted. Six weak learners including conjunctive rules, decision stump, decision table, pruned model trees, rule model trees, and multilayer perceptron, implemented in the data mining system WEKA, were applied. All algorithms were employed to regression problems of property valuation using real-world data derived from the cadastral system and cleansed by property valuation experts.

The lowest prediction error expected by Base test set and the highest error by 30%H test set could be observed only in the case of model trees and a neural network. Model trees and decision tables revealed significantly better performance than conjunctive rule and decision stump.

It is planned to explore resampling methods ensuring faster data processing such as random subspaces, subsampling, and techniques of determining the optimal sizes of multi-model solutions which lead to achieve both low prediction error and an appropriate balance between accuracy and complexity.

Acknowledgments. This paper was partially supported by Ministry of Science and Higher Education of Poland under grant no. N N519 407437.

References

1. Bańczyk, K.: Multi-agent system based on heterogeneous ensemble machine learning models. Master's Thesis, Wrocław University of Technology, Wrocław, Poland (2011)
2. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
3. Büchlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30, 927–961 (2002)
4. Cordón, O., Quirin, A.: Comparing Two Genetic Overproduce-and-choose Strategies for Fuzzy Rule-based Multiclassification Systems Generated by Bagging and Mutual Information-based Feature Selection. Int. J. Hybrid Intel. Systems 7(1), 45–64 (2010)
5. Cunningham, S.J., Frank, E., Hall, M., Holmes, G., Trigg, L., Witten, I.H.: WEKA: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann. New Zealand (2005)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
7. Efron, B., Tibshirani, R.J.: Improvements on cross-validation: the.632+ bootstrap method. Journal of the American Statistical Association 92(438), 548–560 (1997)
8. Friedman, J.H., Hall, P.: On bagging and nonlinear estimation. Journal of Statistical Planning and Inference 137(3), 669–683 (2007)
9. Fumera, G., Roli, F., Serrau, A.: A theoretical analysis of bagging as a linear combination of classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7), 1293–1299 (2008)
10. García, S., Fernandez, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180, 2044–2064 (2010)
11. García, S., Fernandez, A., Luengo, J., Herrera, F.: A Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability. Soft Computing 13(10), 959–977 (2009)
12. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
13. Graczyk, M., Lasota, T., Trawiński, B.: Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 800–812. Springer, Heidelberg (2009)
14. Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K.: Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In: Nguyen, N.T., Le, M.T., Świątek, J., et al. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 340–350. Springer, Heidelberg (2010)
15. Król, D., Lasota, T., Trawiński, B., Trawiński, K.: Investigation of Evolutionary Optimization Methods of TSK Fuzzy Model for Real Estate Appraisal. International Journal of Hybrid Intelligent Systems 5(3), 111–128 (2008)
16. Krzystanek, M., Lasota, T., Telec, Z., Trawiński, B.: Analysis of Bagging Ensembles of Fuzzy Models for Premises Valuation. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 330–339. Springer, Heidelberg (2010)

17. Lasota, T., Mazurkiewicz, J., Trawiński, B., Trawiński, K.: Comparison of Data Driven Models for the Validation of Residential Premises using KEEL. International Journal of Hybrid Intelligent Systems 7(1), 3–16 (2010)
18. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Exploration of Bagging Ensembles Comprising Genetic Fuzzy Models to Assist with Real Estate Appraisals. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 554–561. Springer, Heidelberg (2009)
19. Polikar, R.: Ensemble Learning. Scholarpedia 4(1), 2776 (2009)
20. Schapire, R.E.: The Strength of Weak Learnability. Mach. Learning 5(2), 197–227 (1990)
21. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Investigation of Bagging Ensembles of Genetic Neural Networks and Fuzzy Systems for Real Estate Appraisal

Olgierd Kempa[1], Tadeusz Lasota[1], Zbigniew Telec[2], and Bogdan Trawiński[2]

[1] Wrocław University of Environmental and Life Sciences,
Dept. of Spatial Management ul. Norwida 25/27, 50-375 Wrocław, Poland
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
olgierd_kempa@vp.pl, tadeusz.lasota@wp.pl,
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl

Abstract. Artificial neural networks are often used to generate real appraisal models utilized in automated valuation systems. Neural networks are widely recognized as weak learners therefore are often used to create ensemble models which provide better prediction accuracy. In the paper the investigation of bagging ensembles combining genetic neural networks as well as genetic fuzzy systems is presented. The study was conducted with a newly developed system in Matlab to generate and test hybrid and multiple models of computational intelligence using different resampling methods. The results of experiments showed that genetic neural network and fuzzy systems ensembles outperformed a pairwise comparison method used by the experts to estimate the values of residential premises over majority of datasets.

Keywords: ensemble models, genetic neural networks, bagging, out-of-bag, property valuation.

1 Introduction

The application of soft computing techniques to assist with real estate appraisals has been intensively studied for last two decades. The main focus has been directed towards neural networks [16], [22], [26], [29], less researchers have been involved in the application of fuzzy systems [1], [11]. So far, we have investigated several approaches to construct predictive models to assist with real estate appraisal encompassing evolutionary fuzzy systems, neural networks, decision trees, and statistical algorithms using MATLAB, KEEL, RapidMiner, and WEKA data mining systems [13], [17], [19]. Quite recently, we have built and tested models employing evolving fuzzy systems eTS [21] and FLEXFIS [23] which treated cadastral data on property sales/purchase transactions as a data stream which in turn could reflect the changes of real estate market in the course of time. We studied also bagging ensemble models created applying such computational intelligence techniques as fuzzy systems, neural networks, support vector machines, regression trees, and statistical regression [14], [18], [20]. The results showed that all ensembles outperformed single base models but one based on support vector machines.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 323–332, 2011.

Bagging ensembles, which besides boosting belong to the most popular multi-model techniques have been focused attention of many researchers for last fifteen years. Bagging, which stands for bootstrap aggregating, devised by Breiman [3] is one of the most intuitive and simplest ensemble algorithms providing a good performance. Diversity of learners is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the original training set. So obtained training data subsets, called also bags, are used then to train different classification or regression models. Finally, individual learners are combined through an algebraic expression, such as minimum, maximum, sum, mean, product, median, etc. [27]. Theoretical analyses and experimental results proved benefits of bagging especially in terms of stability improvement and variance reduction of learners for both classification and regression problems [4], [9], [10].

This collection of methods combines the output of the machine learning systems, in literature called "weak learners" in due to its performance [28], from the group of learners in order to get smaller prediction errors (in regression) or lower error rates (in classification). The individual estimator must provide different patterns of generalization, thus the diversity plays a crucial role in the training process. Otherwise, the ensemble, called also committee, would end up having the same predictor and provide as good accuracy as the single one. It was proved that the ensemble performs better when each individual machine learning system is accurate and makes error on the different instances at the same time.

The goal of the study presented in this paper was twofold. Firstly, we would like to investigate the usefulness of genetic neural networks and genetic fuzzy systems to create ensemble models providing better performance than their single base models. Secondly, we aimed to compared soft computing ensemble methods with a property valuating method employed by professional appraisers in reality. The algorithms were applied to real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions obtained from a cadastral system. The investigation was conducted with a newly developed system in Matlab to generate and test hybrid and multiple models of computational intelligence using different resampling methods. The experiments allowed also for the comparison of different approaches to create bagging ensembles with commonly used 10-fold cross validation as well as with models providing the estimation of the resubstitution error.

2 Methods Used and Experimental Setup

The investigation was conducted with our new experimental system implemented in Matlab environment using Neural Network, Fuzzy Logic, Global Optimization, and Statistics toolboxes [7], [12]. The system was designed to carry out research into machine learning algorithms using various resampling methods and constructing and evaluating ensemble models for regression problems. The main modules of the

system are: data management, experiment design, experiment execution, and result analysis and visualization. At the first stage of development particular emphasis was placed on evolutionary neural networks and fuzzy systems. Partitioning methods implemented so far comprise repeated holdout, repeated cross-validation and multiple bootstrap where percentage of base dataset instances drawn to training sets can be determined by a user while defining a data project. It is planned to extend our experimental system to include random subspaces and other diversity creation methods based on techniques of feature selection.

2.1 Expert's Valuation Method

In order to compare evolutionary machine learning algorithms with techniques applied to property valuation we asked professional appraisers to evaluate premises using historical data of sales/purchase transactions obtained from a cadastral system. In this section the pairwise comparison method used by the experts to estimate the values of premises comprised in our dataset is described. The experts simulated professional appraisers' work in the way it is done in reality.

First of all the whole area of the city was divided into 6 quality zones: 1 - the central one, 2 - near-medium, 3 - eastern-medium, and 4 – western-medium zones, and finally 5 - south-western-distant and 6 - north-eastern-distant zones. Next, the premises located in each zone were classified into 243 groups determined by 5 following quantitative features selected as the main price drivers: *Area, Year, Storeys, Rooms,* and *Centre.* Domains of each feature were split into three brackets as follows.

Area denotes the usable area of premises and comprises small flats up to 40 m^2, medium flats in the bracket 40 to 60 m^2, and big flats above 60 m^2.

Year (*Age*) means the year of a building construction and consists of old buildings constructed before 1945, medium age ones built in the period 1945 to 1960, and new buildings constructed between 1960 and 1996, the buildings falling into individual ranges are treated as in bad, medium, and good physical condition respectively.

Storeys are intended for the height of a building and are composed of low houses up to three storeys, multi-family houses from 4 to 5 storeys, and tower blocks above 5 storeys.

Rooms are designated for the number of rooms in a flat including a kitchen. The data contain small flats up to 2 rooms, medium flats in the bracket 3 to 4, and big flats above 4 rooms.

Centre stands for the distance from the city centre and includes buildings located near the centre i.e. up to 1.5 km, in a medium distance from the centre - in the brackets 1.5 to 5 km, and far from the centre - above 5 km.

Then the prices of premises were updated according to the trends of the value changes over 11 years. Starting from the beginning of 1998 the prices were updated for the last day of subsequent years. The trends were modelled by polynomials of degree three. The chart illustrating the change trend of average transactional prices per square metre is given in Fig. 1.

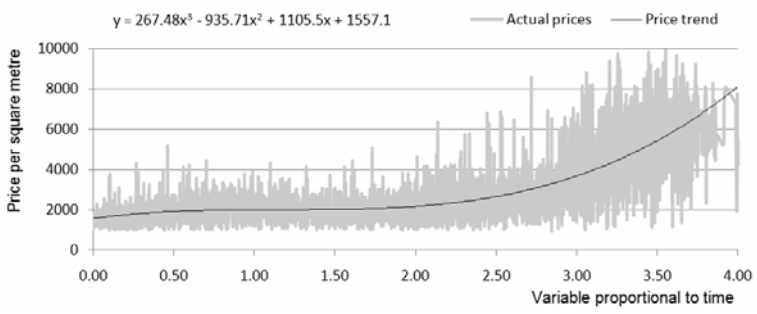


Fig. 1. Change trend of average transactional prices per square metre over time

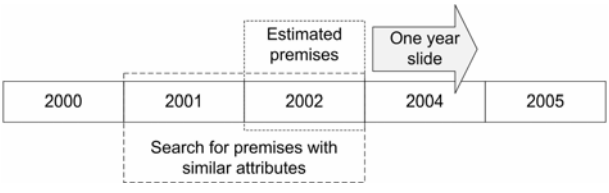


Fig. 2. Time windows used in the pairwise comparison method of experts' estimation

Premises estimation procedure employed a two-year time window to take transaction data of similar premises into consideration (Fig. 2).

1. Take next premises to estimate.
2. Check the completeness of values of all five features and note a transaction date.
3. Select all premises sold earlier than the one being appraised, within current and one preceding year and assigned to the same group.
4. If there are at least three such premises calculate the average price taking the prices updated for the last day of a given transaction year.
5. Return this average as the estimated value of the premises.
6. Repeat steps 1 to 5 for all premises to be appraised.
7. For all premises not satisfying the condition determined in step 4 extend the quality zones by merging 1 & 2, 3 & 4, and 5 & 6 zones. Moreover, extend the time window to include current and two preceding years.
8. Repeat steps 1 to 5 for all remaining premises.

2.2 Genetic Neural Networks and Genetic Fuzzy Systems

In our experiments we employed two basic evolutionary approaches to real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions obtained from a cadastral system, namely genetic neural networks (GNN) and genetic fuzzy systems (GFS). In both techniques we used the same the input and output variables as did the experts in their pairwise comparison

method described above, namely five inputs: *Area, Year, Storeys, Rooms, Centre*, and *Price* as the target variable. The parameters of the architecture of GNN and GFS as well as genetic algorithms are listed in Table 1.

Table 1. Parameters of GNN and GFS used in experiments

GNN	GFS
Network type: feedforward backpropagation	Type of fuzzy system: Mamdani
No. of input variables: 5	No. of input variables: 5
No. of neurons in input layer: 5	No. of input membership functions (mf): 3
No. of hidden layers: 1	No. of output mf: 3
No. of neurons in hidden layer: 5	Type of mf: triangular and trapezoidal
	No. of rules: 15
	Mf parameter variability intervals: ±40%
Chromosome: weights of neuron connections	Chromosome: rules and mf
Chromosome coding: real-valued	Chromosome coding: real-valued
Population size: 150	Population size: 50
Creation function: uniform	Creation function: uniform
Selection function: tournament	Selection function: tournament
Tournament size: 4	Tournament size: 4
Elite count: 2	Elite count: 2
Crossover fraction: 0.8	Crossover fraction: 0.8
Crossover function: two point	Crossover function: two point
Mutation function: Gaussian	Mutation function: custom
No. of generations: 300	No. of generations: 100

Our GNN approach consisted in the evolution of connection weights with a predefined architecture of feedforward network with backpropagation comprising five neurons in an input layer and also five neurons in one hidden layer. Our preliminary tests showed that we can use such a small number of neurons in one hidden layer without the loss of prediction accuracy.

A whole set of weights in a chromosome was represented by real numbers. Similar solutions are described in [15], [30]. In turn, in GFS approach for each input and output variable three triangular and trapezoidal membership functions were automatically determined by the symmetric division of the individual attribute domains. The evolutionary optimization process combined both learning the rule base and tuning the membership functions using real-coded chromosomes. Similar designs are described in [5], [6], [17].

2.3 Dataset Used in Experiments

Real-world dataset used in experiments was drawn from a rough dataset containing above 50 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within eleven years from 1998 to 2008. In this period most transactions were made with non-market prices when the council was selling flats to their current tenants on preferential terms. First of all, transactional records referring to residential premises sold at market prices were selected. Then the dataset was confined to sales transaction data of apartments built before 1997 and where the land was leased on terms of perpetual usufruct. Hence, the final dataset counted 5303 records and comprised all premises which values could be estimated by the experts.

Due to the fact that the prices of premises change substantially in the course of time, the whole 11-year dataset cannot be used to create data-driven models. Therefore it was split into subsets covering individual years, and we might assume that within one year the prices of premises with similar attributes were roughly comparable. The sizes of one-year data subsets are given in Table 1.

Table 2. Number of instances in one-year datasets

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
269	477	329	463	530	653	546	580	677	575	204

2.4 Subsampling and Other Methods Used for Comparison

A series of machine learning experiments was conducted over one-year datasets as base original datasets separately to obtain three bagging ensembles, models providing the resubstitution errors, and 10-fold cross-validation models. Moreover, the property valuation experts built their models using the pairwise comparison method. As a result following models were created and evaluated.

BaseBase – models learned and tested using the same original base dataset, no resampling was used, their performance MSE(BaseBase) refers to the so called resubstitution error or apparent error. The resubstitution errors are overly optimistic because the same samples are used to build and to test the models. They were used to calculate correcting components in the 0.632 bootstrap method.

BagBase – models learned using bags, i.e. bootstrap replicates and tested with the original base datasets, it represents a classic bagging ensemble devised by Breiman. The overall performance of the model MSE(BagBase) equals the average of all MSE values obtained with test set for individual component models. The prediction errors are underestimated since the learning and test sets overlap.

BagOoB – models learned using bags, i.e. bootstrap replicates and tested with the out-of-bag datasets. The overall performance of the model was calculated as the average of all MSE values obtained with test set for individual component models. Due to the fact that training sets comprise on average 63.2 percent of all observations the prediction errors tend to be overestimated.

.632 – model represents the 0.632 bootstrap method correcting the out-of-bag prediction error using the weighted average of the BagOoB and BaseBase MSEs with the weights equal to 0.632 and 0.368 respectively [8].

10cv – ten-fold cross validation, the widely used form of cv for obtaining a reliable estimate of the prediction error. It is known that by growing the number of folds when using k-fold cv the bias can be reduced but at the same time the variance is increased.

Expert – a model based on pairwise comparison approach developed by professional property valuators, its performance was expressed in terms of mean squared error of predicted and actual prices of residential premises.

In the case of bagging method 50 bootstrap replicates (bags) were created on the basis of each base dataset with the number of instances equal to the cardinality of a given dataset. As performance functions the mean square error (MSE) was used and as aggregation functions averages were employed.

3 Results of Experiments

The performance of six models, i.e. *BaseBase*, *BagBase*, *BagOoB*, *.632*, *10cv*, and *Expert* created by GNN and GFS in terms of MSE was presented in Tables 3 and 4 and compared graphically in Figures 3 and 4 respectively. The analysis of the results is carried out separately for GNN and GFS because the experimental conditions did not allow a fair comparison. It can be easily seen that the performance of the experts' method fluctuates strongly achieving for some datasets, i.e. 1999, 2001, and 2008, excessively high MSE values. The differences between *BagBase* and *BagOoB* are apparent in favour of the former method. 10-fold cross validation which is often employed as a method of evaluating single base models reveals worse predictive accuracy than a classic bagging method with a whole original dataset as a test set, i.e. *BagBase*. .632 - the corrected bagging technique provides better results than 10cv for majority of datasets.

Table 3. Performance of models generated using genetic neural networks (GNN)

Dataset	BaseBase	BagBase	BagOoB	.632	10cv	Expert
1998	0.01039	0.01085	0.01274	0.01188	0.01237	0.01262
1999	0.00785	0.00863	0.00930	0.00876	0.00875	0.01335
2000	0.00679	0.00709	0.00787	0.00747	0.00829	0.01069
2001	0.00434	0.00461	0.00499	0.00475	0.00712	0.01759
2002	0.00524	0.00544	0.00610	0.00578	0.00579	0.00760
2003	0.00594	0.00630	0.00650	0.00629	0.00670	0.00753
2004	0.00904	0.01053	0.01094	0.01024	0.01159	0.01049
2005	0.00610	0.00651	0.00673	0.00650	0.00682	0.00936
2006	0.00889	0.00897	0.00954	0.00930	0.00915	0.00905
2007	0.00426	0.00429	0.00464	0.00450	0.00518	0.00659
2008	0.00550	0.00576	0.00668	0.00624	0.00670	0.01400

Table 4. Performance of models generated using genetic fuzzy systems (GFS)

GFS	BaseBase	BagBase	BagOoB	.632	10cv	Expert
1998	0.01003	0.01106	0.01380	0.01241	0.01517	0.01262
1999	0.00902	0.00980	0.01099	0.01027	0.01081	0.01335
2000	0.00662	0.00834	0.01072	0.00921	0.00861	0.01069
2001	0.00483	0.00552	0.00647	0.00587	0.00623	0.01759
2002	0.00537	0.00637	0.00710	0.00647	0.00654	0.00760
2003	0.00662	0.00741	0.00791	0.00743	0.00788	0.00753
2004	0.01170	0.01079	0.01207	0.01194	0.01087	0.01049
2005	0.00666	0.00725	0.00826	0.00767	0.00811	0.00936
2006	0.01108	0.01031	0.01162	0.01142	0.01092	0.00905
2007	0.00587	0.00519	0.00605	0.00598	0.00621	0.00659
2008	0.00709	0.00713	0.00948	0.00860	0.00939	0.01400

The Friedman test performed in respect of MSE values of all models built over eleven one-year datasets showed that there are significant differences between some models in the case both genetic neural networks and genetic fuzzy sets. Average ranks of individual models are shown in Table 5, where the lower rank value the better model. In Table 6 the results of nonparametric Wilcoxon signed-rank test to pairwise comparison of the model performance are presented. The zero hypothesis stated there were not significant differences in accuracy, in terms of MSE, between given pairs of

models. In Table 6 + denotes that the model in the row performed significantly better than, – significantly worse than, and ≈ statistically equivalent to the one in the corresponding column, respectively. In turn, / (slashes) separate the results for GNN and GFS respectively. The significance level considered for the null hypothesis rejection was 5%. Main outcome is as follows: BaseBase models showed significantly better performance than any other model but one BagBase for GFS, BagBase models performed significantly better than any other model but one BaseBase which turned out to be statistically equivalent. These two observations conform with theoretical considerations. Expert models revealed significantly lower MSE than most of other models except BagOoB, .632, and 10cv for GFS.

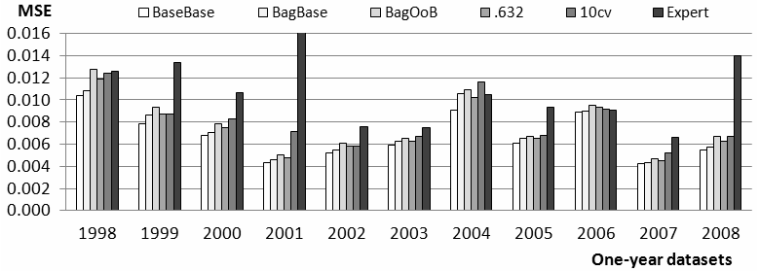


Fig. 3. Performance of models generated using genetic neural networks (GNN)

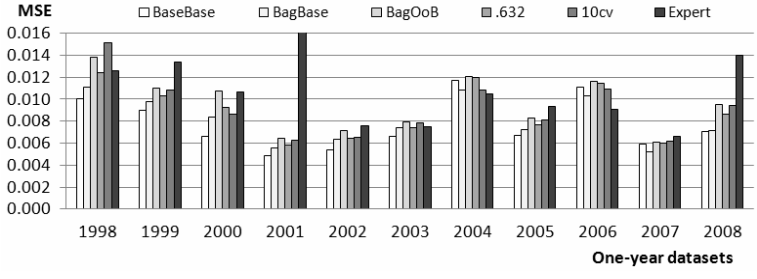


Fig. 4. Performance of models generated using genetic fuzzy systems (GFS)

Table 5. Average rank positions of models determined during Friedman test

	1st	2nd	3rd	4th	5th	6th
GNN	BaseBase (1.00)	BagBase (2.36)	.632 (3.00)	BagOoB (4.64)	10cv (4.64)	Expert (5.36)
GFS	BaseBase (1.64)	BagBase (1.91)	.632 (3.45)	10cv (4.09)	Expert (4.64)	BagOoB (5.27)

Table 6. Results of Wilcoxon tests for the performance of GNN/GFS models

	BaseBase	BagBase	BagOoB	.632	10cv	Expert
BaseBase		+ / ≈	+ / +	+ / +	+ / +	+ / +
BagBase	– / ≈		+ / +	+ / +	+ / +	+ / +
BagOoB	– / –	– / –		– / –	≈ / ≈	+ / ≈
.632	– / –	– / –	+ / +		+ / ≈	+ / ≈
10cv	– / –	– / –	≈ / ≈	– / ≈		+ / ≈
Expert	– / –	– / –	– / ≈	– / ≈	– / ≈	

4 Conclusions and Future Work

The experiments, aimed to compare the performance of bagging ensembles built using genetic neural networks and genetic fuzzy systems were conducted. Moreover, the predictive accuracy of a pairwise comparison method applied by professional appraisers in reality was compared with soft computing machine learning models for residential premises valuation. The investigation was carried out with our new experimental system implemented in Matlab designed to conduct research into machine learning algorithms using various resampling methods and constructing and evaluating ensemble models for regression problems.

The overall results of our investigation were as follows. The bagging ensembles created using genetic neural networks as well as genetic fuzzy systems revealed prediction accuracy not worse than the experts' method employed in reality. It confirms that automated valuation models can be successfully utilized to support appraisers' work. Moreover, the bagging ensembles outperformed single base models assessed using 10-fold cross validation.

Due to excessive times of generating ensemble models on the basis of both genetic neural networks and genetic fuzzy systems it is planned to explore resampling methods ensuring faster data processing such as random subspaces, subsampling, and techniques of determining the optimal sizes of multi-model solutions. This can lead to achieve both low prediction error and an appropriate balance between accuracy and complexity as shown in recent studies [2], [24], [25].

Acknowledgments. This paper was partially supported by Ministry of Science and Higher Education of Poland under grant no. N N519 407437.

References

1. Bagnoli, C., Smith, H.C.: The Theory of Fuzzy Logic and its Application to Real Estate Valuation. Journal of Real Estate Research 16(2), 169–199 (1998)
2. Borra, S., Di Ciaccio, A.: Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics & Data Analysis 54(12), 2976–2989 (2010)
3. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 123–140 (1996)
4. Büchlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30, 927–961 (2002)
5. Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., Magdalena, L.: Ten years of genetic fuzzy systems: current framework and new trends. Fuzzy Sets and Systems 141, 5–31 (2004)
6. Cordón, O., Herrera, F.: A Two-Stage Evolutionary Process for Designing TSK Fuzzy Rule-Based Systems. IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics 29(6), 703–715 (1999)
7. Czuczwara, K.: Comparative analysis of selected evolutionary algorithms for optimization of neural network architectures. Master's Thesis, Wrocław University of Technology, Wrocław, Poland (2010) (in Polish)
8. Efron, B., Tibshirani, R.J.: Improvements on cross-validation: the.632+ bootstrap method. Journal of the American Statistical Association 92(438), 548–560 (1997)
9. Friedman, J.H., Hall, P.: On bagging and nonlinear estimation. Journal of Statistical Planning and Inference 137(3), 669–683 (2007)

10. Fumera, G., Roli, F., Serrau, A.: A theoretical analysis of bagging as a linear combination of classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7), 1293–1299 (2008)
11. González, M.A.S., Formoso, C.T.: Mass appraisal with genetic fuzzy rule-based systems. Property Management 24(1), 20–30 (2006)
12. Góral, M.: Comparative analysis of selected evolutionary algorithms for optimization of fuzzy models for real estate appraisals. Master's Thesis, Wrocław University of Technology, Wrocław, Poland (2010) (in Polish)
13. Graczyk, M., Lasota, T., Trawiński, B.: Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 800–812. Springer, Heidelberg (2009)
14. Graczyk, M., Lasota, T., Trawiński, B., Trawiński, K.: Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In: Nguyen, N.T., Le, M.T., Świątek, J., et al. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 340–350. Springer, Heidelberg (2010)
15. Kim, D., Kim, H., Chung, D.: A Modified Genetic Algorithm for Fast Training Neural Networks. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3496, pp. 660–665. Springer, Heidelberg (2005)
16. Kontrimas, V., Verikas, A.: The mass appraisal of the real estate by computational intelligence. Applied Soft Computing 11(1), 443–448 (2011)
17. Król, D., Lasota, T., Trawiński, B., Trawiński, K.: Investigation of Evolutionary Optimization Methods of TSK Fuzzy Model for Real Estate Appraisal. International Journal of Hybrid Intelligent Systems 5(3), 111–128 (2008)
18. Krzystanek, M., Lasota, T., Telec, Z., Trawiński, B.: Analysis of Bagging Ensembles of Fuzzy Models for Premises Valuation. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 330–339. Springer, Heidelberg (2010)
19. Lasota, T., Mazurkiewicz, J., Trawiński, B., Trawiński, K.: Comparison of Data Driven Models for the Validation of Residential Premises using KEEL. International Journal of Hybrid Intelligent Systems 7(1), 3–16 (2010)
20. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Exploration of Bagging Ensembles Comprising Genetic Fuzzy Models to Assist with Real Estate Appraisals. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 554–561. Springer, Heidelberg (2009)
21. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Investigation of the eTS Evolving Fuzzy Systems Applied to Real Estate Appraisal. Journal of Multiple-Valued Logic and Soft Computing (2011) (in print)
22. Lewis, O.M., Ware, J.A., Jenkins, D.: A novel neural network technique for the valuation of residential property. Neural Computing & Applications 5(4), 224–229 (1997)
23. Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O., Lasota, T.: On Employing Fuzzy Modeling Algorithms for the Valuation of Residential Premises (2011) (to be published)
24. Martínez-Muñoz, G., Suárez, A.: Out-of-bag estimation of the optimal sample size in bagging. Pattern Recognition 43, 143–152 (2010)
25. Molinaro, A.N., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
26. Peterson, S., Flangan, A.B.: Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research 31(2), 147–164 (2009)
27. Polikar, R.: Ensemble Learning. Scholarpedia 4(1), 2776 (2009)
28. Schapire, R.E.: The Strength of Weak Learnability. Mach. Learning 5(2), 197–227 (1990)
29. Worzala, E., Lenk, M., Silva, A.: An Exploration of Neural Networks and Its Application to Real Estate Valuation. The Journal of Real Estate Research 10(2), 185–201 (1995)
30. Yao, X.: Evolving artificial neural networks. Proc. of the IEEE 87(9), 1423–1444 (1999)

Multiple Classifier Method for Structured Output Prediction Based on Error Correcting Output Codes

Tomasz Kajdanowicz[1], Michal Wozniak[2], and Przemyslaw Kazienko[1]

[1] Wroclaw University of Technology, Wroclaw, Poland
Faculty of Computer Science and Management
{tomasz.kajdanowicz,kazienko}@pwr.wroc.pl
[2] Wroclaw University of Technology, Wroclaw, Poland
Faculty of Electronics
michal.wozniak@pwr.wroc.pl

Abstract. It is proposed in the paper a new method for structured output prediction using ensemble of classifiers composed on the basis of Error-Correcting Output Codes. It was presented that newly presented Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes requires comparable computation time in comparison to other accurate algorithms and simultaneously in the classification of more complex structures it provides much better results in the same computation time.

Keywords: Structured Output Prediction, Classifier Ensembles, Multiple Classifier Systems, Error-Correcting Output Codes (ECOC).

1 Introduction

Error-Correcting Output Codes (ECOC) are used to address diverse problems in pattern recognition, e.g. in designing of combined classifiers ECOC provide diversity between classifiers by means of dichotomies, what may result in accurate classification. Its ability to reinforce the classification accuracy may be utilized in the complex problem concerning classification of complex outputs, especially structures - structured output prediction. The main motivation of this paper is to propose the new accurate method for Structured Output Prediction making use of ECOC and the concept of Multiple Classifier Systems (MCS).

1.1 Multiple Classifier Systems

Problems of designing combined classifiers, known as Multiple Classifier Systems (MCSs) of Classifier Ensembles, could be grouped into three main issues:

1. how to choose the topology of such a recognition system - a parallel topology usually is chosen because it is intuitive and well researched;

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 333–342, 2011.

2. how to nominate classifier to the ensemble to ensure high diversity within the group of the chosen individual classifiers;
3. what method of fusion should be proposed to exploit strengths of the chosen individual classifiers.

Focusing on the second, above mentioned problem, the strategy for generating the ensemble members has to improve its diversity. There can be used various components of MCS to enforce classifier diversity:

- using different input data, e.g. using various partitions of the dataset or generating different datasets by data splitting, cross-validated committee, bagging, boosting [11], hoping that the classifiers trained on the different inputs appear to be complementary;
- using classifiers with the same input and output, but trained on the basis of different models or model's versions.
- using classifiers with different outputs i.e., each individual classifier could be trained to solve subset of multi-class problem (e.g., binary classifier - one class against the rest strategy) and fusion method should recover the whole set of predefined classes. This is a well known technique called Error-Correcting Output Codes (ECOC) [4].

1.2 Error-Correcting Output Codes (ECOC)

Originally, Error-Correcting Output Codes (ECOC) was developed in the field of pattern recognition for problems with multiple classes. The idea was to avoid solving the multi-class problem directly but to decompose it into dichotomies instead. Therefore, a multi-class pattern recognition problem can be decomposed in the finite quantity of the two-class classification problems [22]. Thus, the aggregated binary classifiers should be able to recognize a native set of predefined classes by dividing pattern recognition problem into dichotomies [8].

Usually, the combination of binary classifiers is made via a simple nearest-neighbor rule, which finds the closest class to the outputs of the binary classifiers (according to a given metric). The most common variations of the binary classifier combinations are: one-against-one and one-against-all [5]. The former produces an intuitive multi-class classifier where at least one binary classifier corresponds to each class. The hypothesis that the given object belongs to the selected class is verified against its membership to one of the others. Such an approach has a flow in case of conflicting answers from classifiers which is not quite straightforward. The second approach - one-against-all method, usually uses Winner Takes All (WTA) rule. Each classifier is trained on instances of the separate class which becomes the first class, all the other classes correspond to the second one. Final classification is made on the basis of support functions using maximum rule. Mentioned above ECOC was proposed as a combination model in [4]. In this approach, each sequence of bits produced by a set of binary classifiers is associated with codewords during learning. The ECOC selects a class with the smallest Hamming distance to its codeword.

1.3 Structured Prediction

According to the proposal from [3], it is assumed that the structured prediction problem is a cost-sensitive classification problem, where each classification output y_i represents a target structure and is coded as a variable-length vector. This vector notation is very useful for encoding. The ith data instance is represented by a sequence of T values (T-length sequence): y_i=(y_i^1, y_i^2, ..., y_i^T), and $y_i^\mu \in C$, where C is a finite set of classes. Additionally, each structure's element $y_i^{\mu_j}$ may be correlated with other elements $y_i^{\mu_k}$ of the structure, i.e. and the appropriate dependency states the profile of structure.

Generally, structured prediction methods can be categorized into two different groups [18]: problem transformation methods, and algorithm adaptation methods. Whereas the former group of methods is independent from algorithms and concerns the transformation of multi-class classification task into one or more single-class classification, the latter adapts existing learning algorithms in order to handle multi-class data directly. This paper focuses on the first group of methods.

As the nature of structured prediction problems is complex, the majority of proposed algorithms is based on the well known binary classification methods adapted in the specific way [14]. The most natural adaptation is structured perceptron [2] that has minimal requirements on the output space shape and is easy to implement. However, it provides somewhat limited generalization accuracy. An example adaptation of the popular backpropagation algorithm is BPMLL [23] where a new error function takes multiple target into account. The next adaptation extends the original AdaBoost algorithm to the sequence labeling case without reducing it to several two-class problems; it is the AdaBoostSeq algorithm proposed in [9,10].

Another solution are Max-margin Markov Nets that consider the structured prediction problem as a quadratic programming problem [16]. They are very useful; unfortunately, they perform very slow. The next, more flexible approach is an alternative adjustment of logistic regression to the structured outputs called Conditional Random Fields [13]. This method provides probabilistic outputs and good generalization, but again, it is relatively slow. Some other, similar to Max-margin Markov Net technique, is Support Vector Machine for Interdependent and Structured Outputs (SVM^{STRUCT}) [17], which applies variety of loss functions.

Some other algorithms from the lazy learning group realizing the structured prediction task are MLkNN and BRkNN [24]. Both of them extend the popular k Nearest Neighbors (kNN) lazy learning algorithm using a Bayesian approach and the maximum aposteriori principle to assign the label (class) set based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors. An alternate, based on meta learning approach, are Hierarchical multi-label classifiers (HMC): HOMER [20], which construct a hierarchy of multi-label classifiers, and RAkEL [19], an ensemble of classifiers trained by means of different small random subset of the label set.

Overall, the structured prediction is a research problem that emerges in many application domains, among others in protein function classification [23], semantic classification of images [1] or text categorization [15].

The main motivation in designing an ensemble method based on error-correcting output codes for structured prediction is to utilize the powerful ECOC concept to encode whole structure with a codewords, build dichotomies and apply multiple classifiers that result in the reasonable accuracy.

2 Multiple Classifier Method for Structured Output Prediction Based on Error Correcting Output Codes

The new Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes (MCSP-ECOC) proposed in this paper consist of six basic groups of operations, see Fig. 1. Primarily, distinct structured targets are extracted from the learning data (training and testing set). It means that the system discovers all structures (sequences) that differs on at least one item in the sequence, e.g. for the 3-length sequence ($T = 3, C = \{0, 1\}$) we can have up to 8 possible distinct targets (0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1). Obviously, in the dataset only some of them may occur. Assume that only the following patterns exist in the learning dataset: (0,0,0), (0,0,1), (1,0,1), (1,1,0).

When all target schemes, which occur in the input data, are discovered each of them is labeled with the temporal single class. For example, we have (0,0,0) - class A, (0,0,1) - class B, (1,0,1) - class C, (1,1,0) - class D. For each class a dichotomy of training data is constructed according to the choice of the dichotomy construction method. For example, for the one-pair-class code generation method, it provides four dichotomies: 'A - not A (the rest - BCD)', 'B - not B (ACD)', 'C - not C', 'D - not D' and the obtained dychotomy map is (by rows): for class A - '1000', class B - '0100', class C - '0010', class D - '0001'. Obviously, there exist numerous methods for code generation, other than one-pair-class code [12].

Afterwords, for each dichotomy a binary classifier is trained. Using the classifiers trained for all dichotomies a testing phase starts with the assignment of binary class to the cases from the test data. Based on the code generated by all binary assignments (classification results) nearest dichotomy code is retrieved, e.g. code '0100' corresponding to the temporal class B (in the implementation Hamming distance was used, see Sec. 3). The appropriate temporal class is assigned to the case from the nearest dichotomy (e.g. temporal class B). In the final step of the method, the temporal class is decomposed into the original target scheme (sequence), e.g. for the temporal class B, we obtain a sequence (0,0,1) in the classification output.

Presenting the MCSP-ECOC approach for structured prediction, its one property needs to be recalled. Structured prediction based on transformation of multi-class classification task into one or more single-class classification strictly depends on the size of the structure and its patterns' diversity. The bigger and

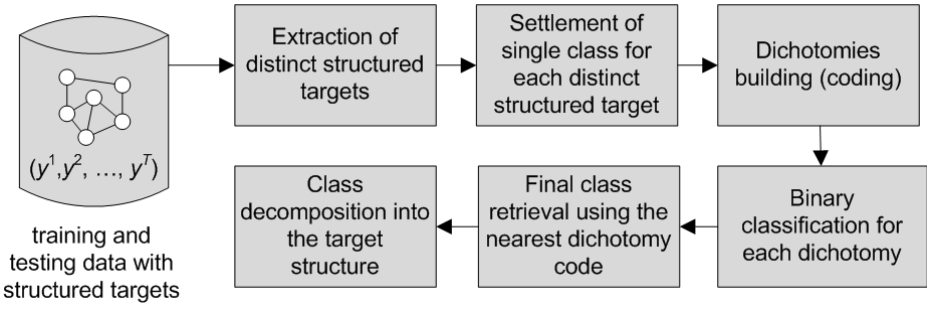


Fig. 1. Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes

more diverse structures the more complex code has too be composed in order to keep its ability of auto-correction. For instance, a sequential structure of a length equals to 10 elements and with 100 distinct patterns requires 100 temporal classes and at least 100 classifiers using one-per-class code (low quality encoding) or as much as 4950 classifiers using all-pairs code (much better encoding) [12].

3 Experiments and Results

The main objective of the performed experiments was to evaluate the classification accuracy of the new proposed method called Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes (MCSP-ECOC). The MCSP-ECOC method was examined according to hamming loss, classification accuracy and computation time for three distinct datasets together with the other state-of-the-art algorithms, representative for the structured prediction problem, namely BPMLL, MLkNN, BRkNN, HMC, HOMER, RAkEL and AdaBoostSeq (ABS).

According to the nature of the structured prediction standard approaches to classification, the typical evaluation my not be sufficient and it requires different evaluation measures than those used in traditional single-label classification. Some standard evaluation measures of multi-class classifiers from the previous work have been used in the experiments. The utilized measures are calculated based on the differences of the actual and the predicted sets of labels (classes) over all cases x_i in the test set. The first measure is Hamming Loss HL, which was proposed in [15] and is defined as:

$$HL = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i \triangle F(x_i)}{|Y_i|} \tag{1}$$

where: N is the total number of cases x_i in the test set; Y_i denotes actual (real) labels (classes) in the sequence, i.e. entire structure corresponding to case

x_i; $F(x_i)$ is a sequence of labels predicted by classifier and $\triangle$ stands for the symmetric difference of two vectors, which is the vector-theoretic equivalent of the exclusive disjunction in Boolean logic.

The second evaluation measure utilized in the experiments is Classification Accuracy CA [7], defined as:

$$CA = \frac{1}{N} \sum_{i=1}^{N} I(Y_i = F(x_i)) \tag{2}$$

where: N, Y_i, $F(x_i)$ have the same meaning as in Eq. 1, $I(true) = 1$ and $I(false) = 0$.

Measure CA is a very strict evaluation measure as it requires the predicted sequence of labels to be an exact match of the true set of labels.

The performance of the analyzed methods was evaluated using 10-fold cross-validation and the evaluation measures from Eq. 1 and Eq. 2. These two metrics are widely-used in the literature and are indicative for the performance of multi-label classification methods. Additionally, the computation time has been monitored.

The experiments were carried out on three datasets from three diverse application domains: semantic scene analysis, bioinformatics and music categorization. The image dataset *scene* [1] semantically indexes still scenes. The biological dataset *yeast* [6] is concerned with protein function classification. The music dataset *emotions* [21], in turn, contains data about songs categorized into one or more classes of emotions.

Table 1. Datasets used in the experiments

	Dataset	Cases	Attributes	Sequence length	Number of patterns	Number of classifiers
1	scene	2407	294	6	15	105
2	yeast	2417	203	14	198	19503
3	emotions	593	72	6	27	351

The basic statistics of utilized datasets, such as the number of cases, the number of numeric and discrete attributes, the length of label sequence, the number of distinct patterns in structured targets as well as the number of classifiers built with all pairs coding technique are presented in Table 1.

The Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes (MCSP-ECOC) was compared to some other methods, in particular to AdaBoostSeq, BPMLL, MLkNN, BRkNN, HMC, HOMER and RAkEL. MCSP-ECOC utilized kNN classifier with $k = 3$ and Hamming distance measure was applied to class assignment based on the nearest dichotomy.

As regards Hamming Loss HL, Eq. 1, the MCSP-ECOC algorithm provided the worst results in comparison with all other tested algorithms. For instance in evaluation based on the *yeast* dataset, MCSP-ECOC provided 60% worse

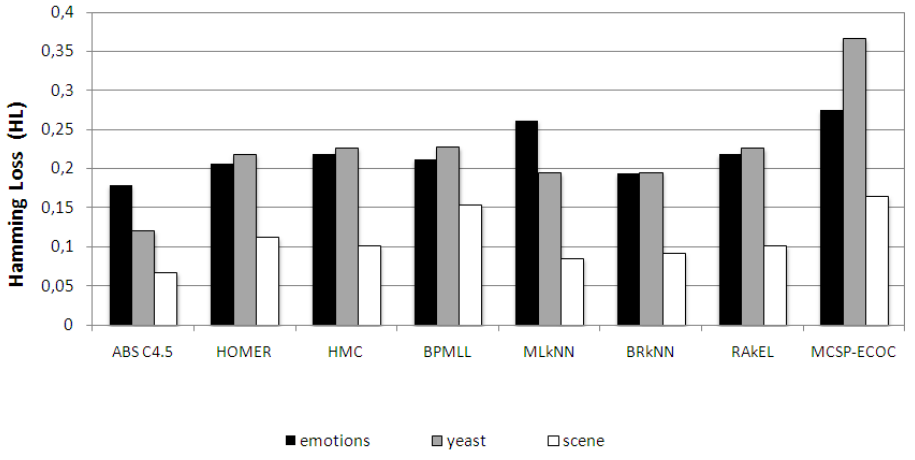


Fig. 2. Hamming loss (HL) measure for examined algorithms on scene, yeast and emotions datasets

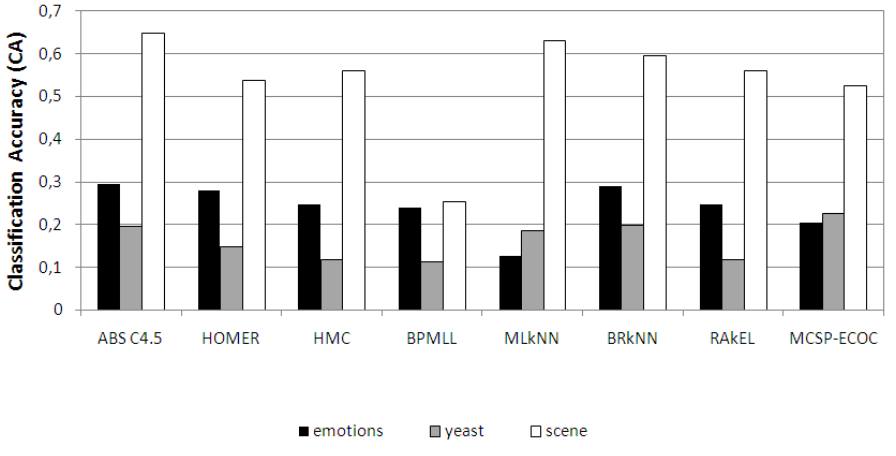


Fig. 3. Classification Accuracy measure for examined algorithms on scene, yeast and emotions datasets

results compared to the second worse BPMLL algorithm, Fig. 2. However, it does not mean that the overall classification accuracy of MCSP-ECOC will result in unacceptable outcome.

It is worth to mention that while the classification accuracy for all other examined algorithms depends directly on hamming loss, MCSP-ECOC has additional feature of code correction. Its high hamming loss does not mean the classification accuracy have to be unsatisfactory.

In fact, while considering the classification accuracy CA, Eq. 2, MCSP-ECOC presents satisfactory results. For *emotions* and *scene* datasets it provided results

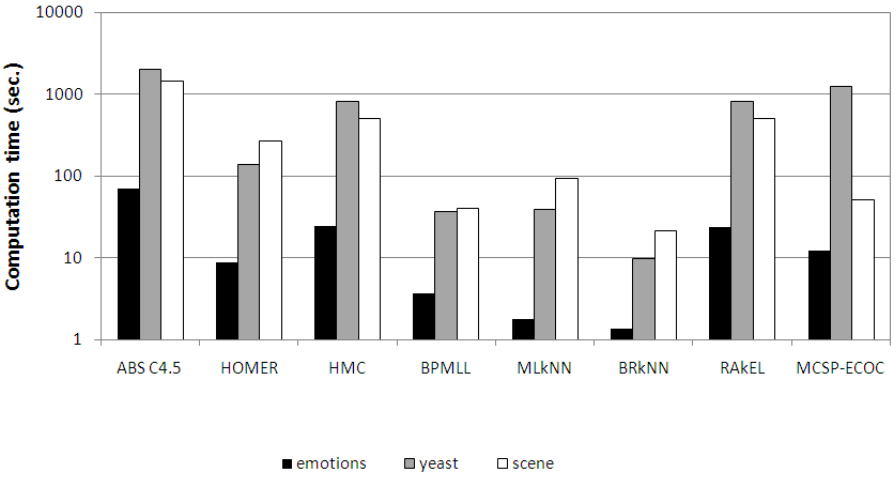


Fig. 4. Execution time in seconds for examined algorithms on scene, yeast and emotions datasets

worse by 17% and 3%, accordingly, in comparison to average result of all other approaches. However for the most complex, *yeast* dataset MCSP-ECOC provided results better by 48% than the average of others and 15% better than the second best result (AdaBoostSeq C4.5), Fig. 3.

Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes method requires comparable computation time to other accurate algorithm, i.e. RAkEL, AdaBoostSeq and HMC4. However, in the classification of more complex structures (*yeast* dataset) it provides in the similar time much better results.

4 Conclusions

In this paper it was proposed a method for structured prediction using ensemble of classifiers composed on the basis of Error-Correcting Output Codes. It was presented that Multiple Classifier Method for Structured Output Prediction based on Error Correcting Output Codes requires comparable computation time to other accurate algorithm, but in the classification of more complex structures it provides in the same time much better results.

Further experiments will be concerning the impact of structural patterns variety on the computation time, the dependency between the coding method and the accuracy of the classification as well as the coding method adequacy to distinct types of structures.

Acknowledgement. This work was supported by The Polish Ministry of Science and Higher Education the research project 2010-2013, and Fellowship co-financed by European Union within European Social Fund.

References

1. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. Pattern Recognition 37, 1757–1771 (2004)
2. Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In: Conference on Empirical Methods in Natural Language Processing 2002, vol. 10, pp. 1–8 (2002)
3. Daume, H., Langford, J., Marcu, D.: Search-based structured prediction. Machine Learning 75, 297–325 (2009)
4. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2, 263–286 (1995)
5. Duan, K., Keerthi, S.S., Chu, W.: Multi-Category Classification by Soft-Max Combination of Binary Classifiers. In: 4th International Workshop, June 11-13, pp. 125–134 (2003)
6. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Information Processing Systems, vol. 14 (2001)
7. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the 3005 ACM Conference on Information and Knowledge Management 2005, pp. 195–200 (2005)
8. Hong, J., Min, J., Cho, U., Cho, S.: Fingerprint classification using one-vs-all support vector machines dynamically ordered with naive Bayes classifiers. Pattern Recognition 41, 662–671 (2008)
9. Kajdanowicz T., Kazienko P.: Boosting-based Sequence Prediction. New Generation Computing (2011)
10. Kajdanowicz, T., Kazienko, P.: Base Classifiers in Boosting-based Classification of Sequential Structures. Neural Network World 20(7), 839–851 (2010)
11. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New Jersey (2004)
12. Kuncheva, L.I.: Using diversity measures for generating error-correcting output codes in classifier ensembles. Pattern Recognition Letters 26, 83–90 (2005)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning ICML 2001, pp. 282–289 (2001)
14. Nguyen, N., Guo, Y.: Comparisons of Sequence Labeling Algorithms and Extensions. In: International Conference on Machine Learning ICML 2000, pp. 681–688 (2007)
15. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
16. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In: Advances in Neural Information Processing Systems, vol. 16, pp. 25–32. MIT Press, Cambridge (2004)
17. Tsochantaridis, I., Hofmann, T., Thorsten, J., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research 6, 1453–1484 (2005)
18. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, 1–13 (2007)
19. Tsoumakas, G., Vlahavas, I.: Random k-Labelsets: An Ensemble Method for Multilabel Classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007)

20. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and Efficient multi-label Classification in Domains with Large Number of Labels. In: ECML/PKDD 2008 Workshop on Mining Multidimensional Data, MMD 2008 (2008)
21. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label Classification of Music into Emotions. In: Proceedings of the 9th International Conference on Music Information Retrieval ISMIR 2008, pp. 325–330 (2008)
22. Zeng, Q., Zhang, L., Xu, Y., Cheng, L., Yan, X., Zu, J., Dai, G.: Designing expert system for in situ Si3N4 toughened based on adaptive neural fuzzy inference system and genetic algorithms. Materials and Design 30, 256–259 (2009)
23. Zhang, M.L., Zhou, Z.H.: Multi-label neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering 18, 1338–1351 (2006)
24. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048 (2007)

Ontology-Based Resource Management for Cloud Computing

Yong Beom Ma, Sung Ho Jang, and Jong Sik Lee

School of Information Engineering, Inha University, #253, YongHyun-Dong, Nam-Ku, 402-751 Incheon, South Korea
{myb112,ho7809}@hanmail.net, jslee@inha.ac.kr

Abstract. Resource management is a challenging issue in cloud computing. This paper aims to allocate requested jobs to cloud resources suitable for cloud user requirements. To achieve the aim, this paper proposes an ontology-based job allocation algorithm for cloud computing to perform inferences based on semantic meanings. We extract resource candidates depending on user requirements and allocate a job to the most suitable candidate for an agreed Service Level Agreement (SLA). The cloud ontology allows the proposed system to define concepts and describe their relations. Hence, we can process complicated queries for searching cloud resources. To evaluate performance of our system, we conducted some experiments compared with the existing resource management algorithms. Experimental results verify that the ontology-based resource management system improves the efficiency of resource management for cloud computing.

Keywords: Cloud Computing, Resource Management, Job Allocation, Ontology.

1 Introduction

With the development of mobile industry and internet technology, there have been recent researches on cloud computing. Cloud computing provides users IT resources integrated into cloud services by using virtualization and distributed processing technologies. Cloud users can use computing resources or services in anywhere, anytime on demand. Infrastructure as a Service (IaaS) is a representative cloud service model which provides IT infrastructures like servers, storages, network as a service [1][2]. In IaaS model, a cloud provider integrates physical resources into a logical resource called a Virtual Machine (VM), and constructs a virtualized environment which consists of several virtual machines interconnected. Cloud users and providers commit to an agreement called SLA [3].

In cloud computing, the aim of resource management is to allocate resources to a job, which is depending on the agreed SLAs and requested from a user. However, available resources of cloud providers are changed dynamically and the requirement of users is various because resources are virtualized in a cloud computing environment. Therefore, a resource management algorithm is essentially based on the agreed SLA by negotiation process between cloud providers and users. There are various resource management algorithms [4][5][6][7] studied by many researchers to manage

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 343–352, 2011.

resources effectively in cloud computing, but the existing algorithms are confined to the economic way. They also focus on how to satisfy cloud users with restricted requirements or how to gain the maximum profits of providers. However, we need to consider more detailed resource information, such as availability and reliability, in order to improve utilization.

In this paper, we propose an ontology-based resource management system to solve problems of existing resource management algorithms in a cloud computing environment. The proposed system uses an ontology-based job allocation algorithm for resource management in a cloud computing environment. We also propose the cloud ontology, which defines cloud computing concepts and describes their relations. For intelligent resource management, we perform an inference depending on some predefined rules.

This paper is organized as follows: Section 2 presents resource management problems in Cloud computing and discusses weak points of existing resource management algorithms. The design of the ontology-based resource management system for Cloud computing and the ontology-based jobs allocation algorithm are describes in Section 3. Section 4 demonstrates effectiveness of the proposed resource management system with experimental results. Finally, conclusions are presented in Section 5.

2 Related Work

Resource management system plays an important role in allocating jobs to the most suitable resources in cloud computing. Several researchers began to study the resource management for cloud computing, and we discuss the works closely related to our work. There are some resource management algorithms for cloud computing. Y. Murata, R. Egawa, M. Higashida, and H. Kobayashi proposed the history of execution time-based resource management algorithm [4]. The algorithm estimates the job-start time of a queuing system and compares the time with the history of the job-execution time. Subsequently, the new job, which can be executed earlier than other jobs, is allocated to an appropriate resource. The history of execution time-based resource management algorithm provides short waiting time and improves utilization. Resource Allocation Strategy based on Market Mechanism (RAS-M) [5] is a profit optimization resource management algorithm. The RAS-M focused on price adjustment to allocate resources to cloud users with the equilibrium state based on a market mechanism. Authors utilize the principle of price and try to maximize profits of resource consumers and providers. The response time optimization-based resource management algorithm [6] was proposed by M. D. de Assuncao, A. di Costanzo, and R. Buyya. In [6], the response time of requested jobs is used for a performance metric. The algorithm aims to improve the response time of requested jobs.

These algorithms provide the resources with the lowest cost or the fastest computing service based on market mechanism. A cloud computing system has numerous components related with each other. However, it is hard to manage effectively not only the components but also resource information and process complicated queries with diverse factors. To solve these problems, this paper proposes a resource management system depending on the ontology-based job allocation algorithm.

3 Ontology-Based Resource Management System

3.1 Design of Ontology-Based Resource Management System

A cloud provider uses virtualization technology in order to integrate dynamically scalable resources such as physical machines, data storages, and network. The cloud provider constructs virtual machines which need to be rapidly elastic on demand. For execution of virtual machines, a cloud system provides virtualized resources as a service and remote on-demand access. A cloud user outsources a job to the cloud provider and just pays for costs depending on service usage. The cloud provider negotiates with the cloud user and reaches a SLA. The SLA includes properties of cloud resources and the requirements of a cloud user, and guarantees Quality of Service (QoS) to the cloud user. In cloud computing, SLAs are used to ensure that service quality is kept and includes QoS parameters such as response time, reliability, availability, and so on [8].

When a cloud user requires processing a job, a cloud provider constructs virtual machines called resource set and provides the virtual machines (VMs) as a service to the cloud user. Each VM has another characteristic related to QoS and begins the service to meet the agreed SLA. The cloud provider needs to dynamically allocate a job to a specific VM in order to meet the SLA and sustain its operations. To realize resource management, we therefore design an architecture for the ontology-based resource management system for Cloud computing as shown in Fig. 1.

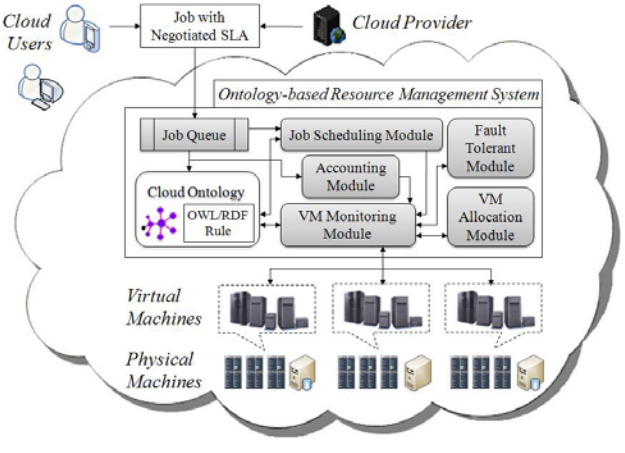


Fig. 1. Architecture of the ontology-based resource management system

Cloud users request a job with a negotiated SLA with a cloud provider and use resources as a service. Cloud users just throw their jobs to the cloud system and then the ontology-based resource management system distributes resources among cloud users. As shown in Fig. 1, the cloud system includes three main parts which are the

ontology-based resource management system, virtual machines, and physical machines. The ontology-based resource management system is composed of seven components; job queue, job scheduling module, cloud ontology with OWL/RDF and Rule, VM allocation module, VM monitoring module, fault tolerant module, and accounting module. The job queue stores jobs requested from users in consecutive order. The submitted SLAs are sent to the cloud ontology and the accounting module. The job scheduling module plays an important role in creating a schedule in order to assign jobs to VMs. The job schedule is created depending on the ontology-based job allocation algorithm described in Section 3.2. The cloud ontology is used to define concepts underlying the proposed system for cloud computing and describe their relations. The VM allocation module allocates VMs to physical resources after jobs are scheduled and assigned to each VM. The VM monitoring module delegates a kind of mediator to monitor the state of virtual machines and manage their resource information. The monitored information is transmitted to cloud ontology for job scheduling. If a failure is detected or a regular job processing is inexecutable, the VM monitoring module instantly informs the fault tolerant module about the failure. The fault tolerant module aims to find an alternative resource and transmit a feedback to the VM monitoring module. The accounting module confirms user requirements with each SLA and bills costs of cloud service usage and shared resources. Virtual machines compose a resource pool from physical machines and perform dynamic provisioning of resources without physical constraints. Resources are shared among multiple jobs and easily duplicated, migrated, and used. Fig. 2 shows an interaction with each other in the ontology-based resource management system in order to assign a job to virtual machines.

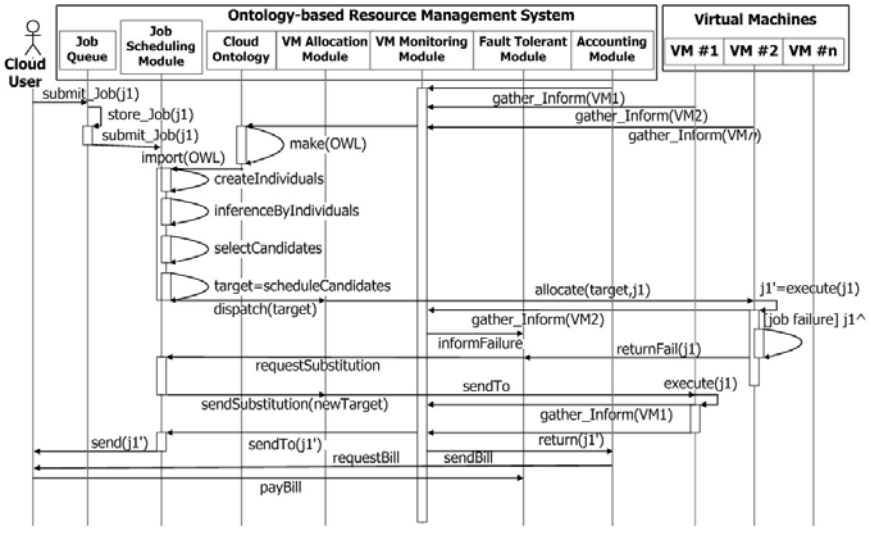


Fig. 2. Job Allocation Process in the ontology-based resource management system

3.2 Design of Cloud Ontology

This section describes an ontology in order to deal with information related to cloud computing resources. The ontology is an explicit specification of a conceptualization [9]. We can use an ontology to give semantic meanings to information corresponding concepts. Well defined ontology can analyze and reuse semantic meanings. If not, we can fall into an infinite loop during the process of inference based on the ontology. In a cloud computing environment, requirements of cloud users are very complex and delicate, resources are composed of various types of physical machines, and jobs need diverse service types. Due to the virtualization of physical resources, resource management is more complicated and difficult. Therefore, we propose the cloud ontology to deal with dynamic resource information for resource management. With the cloud ontology, we can deal with various types of queries and allocate resources suitable for the type of services and the description of jobs requested from cloud users.

Resources are conceptualized to classify the type of virtualized resources because virtualized resources provide physical resources to cloud users in order to process a job in a cloud computing environment. The conceptualization of resources is also required to classify the type of physical resources. Cloud services are classified by the type of jobs requested from cloud users because a job in cloud computing is a form of cloud services. In order to allocate resources to cloud users, we need to consider not only the type of resources but also that of cloud services requested from the cloud user. We use the Ontology Web Language (OWL) [10] to describe classes, constraints, and properties of the cloud ontology in this paper. The OWL allows us to extend concepts related to a specific domain as well as describe the concepts hierarchically.

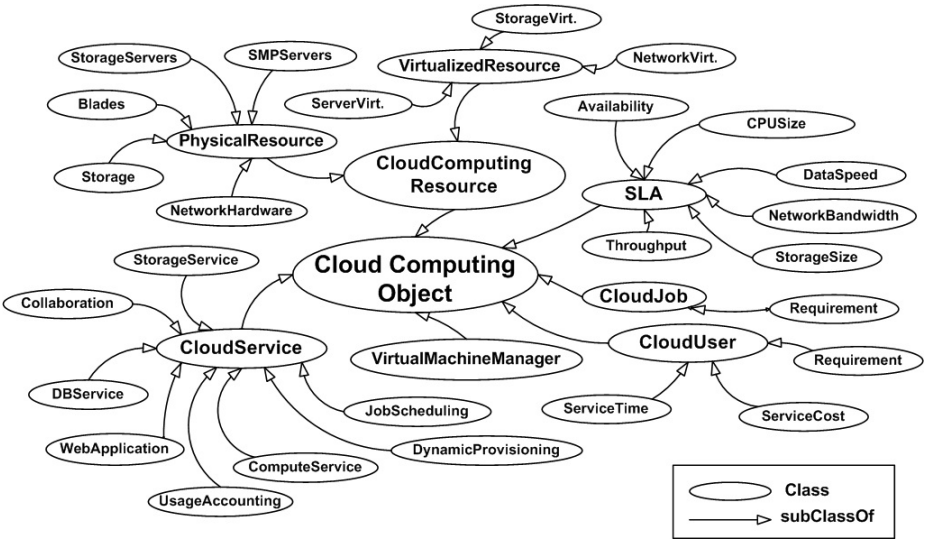


Fig. 3. Overview of cloud ontology

To perform intelligent resource management, we derive basic concepts of the cloud ontology from the ontology-based resource management system in section 3.1. Some core classes are defined based on the proposed system and related to other classes with constraints. To connect pairs of classes and represent a relationship among classes, we use object properties in our ontology. As shown in Fig. 3, we define the CloudComputingResource class, which is one of core classes and includes two general classes; PhysicalResource class and VirtualizedResource. Physical resources should be registered to the virtualized resource for constructing a virtual machine. Virtualized resources should be registered to the virtual machine manger for identification and classified into three types by a service type. Therefore, the PhysicalResource class has two constraints; hasID, and registeredToVR. The VirtualizedResource class has three constraints; hasID, hasServiceType, and registeredToVMM. In cloud computing, jobs should be owned by distinct cloud users and allocated to virtual machines. An agreed SLA is included in a job. Therefore, the CloudJob class has four constraints; hasID, hasOwner, hasAgreedSLA, and allocatedToVR. In this way, object properties can be represented by constraints of classes. We use data properties to specify classes and represent the data range of classes. For example, a physical resource has an ID, the type of an operation system, the size of CPU, and so on. The ID means an identification number and the ID value is generally represented by an integer. The type of an operating system can be represented by string in this way. Some classes have common data properties. The VirtualizedResource class has a property osType, which describes the type of operating system of a virtualized resource. The PhysicalResource class also has a property osType. Therefore, the data property osType can be defined in the CloudComputingResource class because the PhysicalResource class and the VirtualizedResource class inherit the data properties of the CloudComputingResource class.

3.3 Ontology-Based Job Allocation Algorithm

This section describes the ontology-based job allocation algorithm. In a cloud computing environment, cloud users request a job with detailed user requirements. User requirements include deadline, budget, needs for resources such as CPU size, type of operating system, storage size, and QoS parameters such as cost of service, response time, reliability, availability, and so on. Existing resource management systems considers few user requirements depending on a simple query. However, the proposed system can handle user requirements based on an agreed SLA including detailed user requirements. We define the cloud ontology and describe data properties for characteristics of cloud computing objects.

For an inference based on the cloud ontology, we need to a kind of rules and use the Semantic Web Rule Language (SWRL) [11]. A rule of the SWRL has a semantic meaning by expressing the relation between an antecedent and a consequent. An antecedent and a consequent can express a class or the property of the class. It can be expressed in the same form as ClassName(?x) or hasProperty(?x, ?y). Multiple antecedent or consequent processed with conjunctive operations. Table 1 shows some

inference rules that are defined to manage resources for cloud computing. We can give semantic meanings to all cloud computing object information by these inference rules. For example, we can identify where a virtualized resource is included in. The rule 2 extracts available virtualized resources from main user requirements. And we can extract virtualized resources which has a service type depending on user requirements. User requirements including the agreed SLA reflect characteristics of virtualized resources. For example, we can know what kind of cloud services a cloud user requests by some inference rules such as the rule 4, 5, and so on. To identify some information, we need to combine a rule with other rules. For example, resource candidates of the requested cloud service can be extracted by the rule 3. The rule 3 can be expressed by the combination of the rule 2 and one of other rules. The combination of rules allows the inference engine to be more flexible.

Table 1. Inference rules of the cloud ontology

No	Domain rules
1	VirtualizedResource(?x) ∧ registeredToVMM(?x, ?y) ∧ VirtualMachineManager(?y) ∧ groupName(?z) → includedInVMM(?x, ?z)
2	VirtualizedResource (?x) ∧ expectTime(?x, ?a) ∧ [?a<=qDeadline] ∧ expectCost(?x, ?b) ∧ [?b<=qBudget] → AvailableResource(?x)
3	AvailableResource(?x) ∧ hasServiceType (?x, ?a) ∧ [?a=qServiceType] → CandidateResource(?x)
4	VirtualizedResource (?x) ∧ osType(?x, ?a) ∧ [?a=windowXP ∨ windowNT] ∧ cpuSpeed(?x, ?b) ∧ [?b>=1.80] ∧ ramSize(?x, ?c) ∧ [?c>=1024] ∧ bandwidth(?x, ?d) ∧ [?d>=100]→ hasServiceType(?x, ComputeService)
5	VirtualizedResource (?x) ∧ osType(?x, ?a) ∧ [?a=Window2000Server ∨ Window2003Server] ∧ responseTime(?x, ?b) ∧ [?b>=3.0] ∧ cost(?x, ?c) ∧ [?c>=200] ∧ storageSize(?x, ?d) ∧ [?d>=100]→ hasServiceType(?x, WebApplication)
6	VirtualizedResource (?x) ∧ osType(?x, ?a) ∧ [?a=Linux] ∧ cpuSpeed(?x, ?b) ∧ [?b>=2.13] ∧ cost(?x, ?c) ∧ [?c>=100] ∧ storageSize(?x, ?d) ∧ [?d>=1000]→ hasServiceType(?x, StorageService)
…	…

4 Performance Evaluation

We implemented a simulation model by applying the discrete event system specification (DEVS) formalism [12] to the ontology-based resource management system. The Protégé [13] was used to create and edit a cloud ontology, and the Bossam Inference Engine [14] was used to implement inference rules in Table 1. To evaluate the Ontology-based Resource Management Algorithm (ORMA), we conducted some experiments to compare the proposed ORMA with the Profit-Optimization-based Resource Management Algorithm (PORMA) [5] and the Response time-Optimization-based Resource Management Algorithm (RORMA) [6].

In our experiments, cloud computing service was provided by a reliable big business enterprise and 100 cloud resources are dispersed in the simulation environment based on cloud computing. Also, we assumed that cloud resources are virtualized machines because we focus on job allocation to virtual machines. We do not consider that each virtual machine splits the job in lots and allocates the split job to physical machines. As mentioned in section 3.1, virtual machines in cloud computing can dynamically allocate resources suitable for user requirements. Hence, we also assumed that cloud resources in this experiment have the ability to process the jobs requested from cloud users. And jobs requested from cloud users are a total of 3000. A cloud user requests a job with an agreed SLA including deadline, budget, and QoS parameters such as response time, reliability, availability, and so on.

4.1 Simulation Results

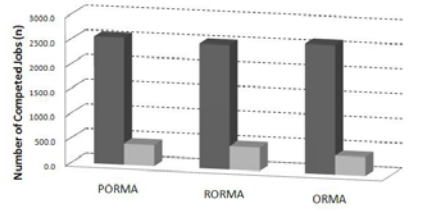
In order to validate the efficiency and excellence of our proposed system, we measured various performance measures; number of completed jobs, throughput, resource utilization, service time. The experimental results are described by the following figures.

Fig. 4 (a) shows the number of completed and lost jobs in this experiment. If there is a job being processed on a cloud resource, a newly requested job will probably wait for completion of the process. The job loss indicates the expired and non-completed jobs. ORMA recorded 2616 completed jobs and 384 lost jobs. PORMA and RORMA recorded less completed jobs than ORMA because the algorithms do not consider diverse QoS parameters like reliability, availability, and so on.

In this experiment, the service time indicates the time needed to complete a job from the generation time of the job. Therefore, the average response time in this experiment is calculated as the subtraction the completion time of the job from the generation time of the job. As mentioned before, the service time can be sum of the processing of a virtual machine and the waiting time of a job. Fig. 4 (b) shows the average service time of each resource management algorithm as explained above. The service time of PORMA is similar to that of ORMA. However, ORMA provided less service time on average since ORMA allocate requested jobs to the suitable cloud resource for satisfying the cloud user requirements.

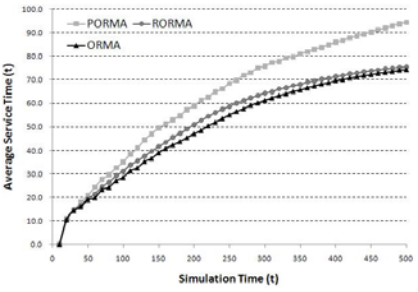
Fig. 4 (c) shows the throughput of three resource management algorithms as time progresses. The throughput can be obtained by the number of completed jobs per time interval. In this experiment, we measured the throughput every 10 simulation time. As shown in Fig. 4 (c), the throughput of ORMA is higher than other resource management algorithms. On average, ORMA provided the throughput of about 5.0 jobs, which is higher than that of other algorithms. That is to say, ORMA can be process more jobs than other algorithms. We also measured the resource utilization.

Fig. 4 (d) shows the average resource utilization of each resource management algorithm with increasing simulation time. ORMA provided the highest resource utilization among three resource management algorithms. On average, ORMA provided resource utilization of about 75.7%, which is 32.3% higher than the utilization provided by PORMA and 13.7% higher than the utilization provided by RORMA. This numerical value shows that ORMA allocates jobs more uniformly to cloud resources than existing resource management algorithms.

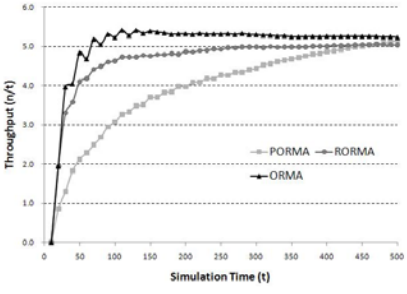


	PORMA	RORMA	ORMA
Completed Jobs	2569.0	2520.0	2616.0
Lost Jobs	431.0	480.0	384.0

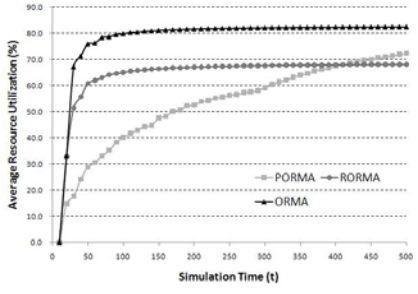
(a) Number of completed and lost jobs



(b) Average service time



(c) Throughput



(d) Average resource utilization

Fig. 4. Performance measures compared with existing resource management algorithms

5 Conclusion

This paper proposes an ontology-based job allocation algorithm for a resource management system in cloud computing environment. Managing virtualized cloud resources and diverse user requirements causes some problems such as complexity of resource information management and difficulty in guaranteeing satisfaction of cloud users. In this paper, we consider virtual machines as the cloud resources of our resource management system instead of physical machines. In order to give various semantic meanings to cloud resource information, our resource management system builds the cloud ontology based on cloud resource information and agreed SLAs. The ontology can be extended by new classes and properties on demand. Therefore, the ontology can be used to process delicate and complicated queries for cloud resources. In addition, we guarantee the quality of service by allocating cloud resources to cloud users dynamically depending on agreed SLAs. To evaluate performances, we performed some experiments to compare the existing resource management algorithms for cloud computing. Experimental results show that the ontology-based job allocation algorithm significantly outperforms existing resource management algorithms for cloud computing.

References

1. Bhardwaj, S., Jain, L., Jain, S.: Cloud Computing: A study of Infrastructure As A Service (IAAS). Journal of Engineering and Information Technology 2(1), 60–63 (2010)
2. Santos, N., Gummadi, K.P., Rodrigues, R.: Towards Trusted Cloud Computing. In: Workshop on Hot Topics in Cloud Computing, USENIX (2009)
3. Patel, P., Ranabahu, A., Sheth, A.: Service Level Agreement in Cloud Computing. In: Cloud Workshops at OOPSLA 2009 (2009), http://knoesis.wright.edu/aboutus/visitors/summer2009/PatelReport.pdf
4. Murata, Y.E., Higashida, R., Kobayashi, M., Cybersci, H.: A History-Based Job Scheduling Mechanism for the Vector Computing Cloud. In: 10th Annual Symposium on Applications & the Internet, pp. 125–128. IEEE Press, New York (2010)
5. You, X., Xu, X., Wan, J., Yu, D.: RAS-M:Resource Allocation Strategy based on Market Mechanism in Cloud Computing. In: 2009 Fourth ChinaGrid Annual Conference, pp. 256–263. IEEE Press, New York (2009)
6. Assuncao, M.D., Costanzo, A.: Evaluating the Cost-Benefit of Using Cloud Computing to Extend the Capacity of Clusters. In: 18th ACM International Symposium on High Performance Distributed Computing, pp. 141–150. ACM Press, New York (2009)
7. Lee, Y.C., Wang, C., Zomaya, A.Y., Zhou, B.B.: Profit-driven Service Request Scheduling in Clouds. In: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 15–24. IEEE Press, New York (2010)
8. Wu, L., Buyya, R.: Service Level Agreement (SLA) in Utility Computing Systems. Thechnical report, CLOUDS-TR-2010-5, Cloud Computing and Distributed Systems Laboratory (2010)
9. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
10. OWL Web Ontology Language Reference, W3C Recommendation, http://www.w3.org/TR/owl-ref/
11. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Recommendation, http://www.w3.org/Submission/SWRL/
12. Ziegler, B.P., Sarjoughian, H.S., Park, S.W., Lee, J.S., Cho, Y.K., Nutaro, J.J.: DEVS modeling and Simulation: a new layer of middleware. In: 3rd Annual International Workshop on Active Middleware Services, pp. 22–31. IEEE Press, New York (2001)
13. Protégé, http://protege.stanford.edu/
14. Jang, M.S., Sohn, J.C.: Bossam: An Extended Rule Engine for OWL Inferencing. In: Antoniou, G., Boley, H. (eds.) RuleML 2004. LNCS, vol. 3323, pp. 128–138. Springer, Heidelberg (2004)

Self-similarity Based Lightweight Intrusion Detection Method for Cloud Computing

Hyukmin Kwon[1], Taesu Kim[1], Song Jin Yu[2], and Huy Kang Kim[1]

[1] Graduate School of Information Security, Korea University, Anam-dong, Seongbuk-gu, Seoul, 136-713, Republic of Korea
{hack,krad,cenda}@korea.ac.kr
[2] Maritime Univ. Division of Shipping Management, Korea Maritime University, 1 Dongsam-Dong, Yeongdo-gu, Busan, 606-791, Republic of Korea
coppers@hhu.ac.kr

Abstract. Information security is the key success factor to provide safe cloud computing services. Despite its usefulness and cost-effectiveness, public cloud computing service is hard to accept because there are many security concerns such as data leakage, unauthorized access from outside the system and abnormal activities from inside the system.

To detect these abnormal activities, intrusion detection system (IDS) require a learning process that can cause system performance degradation. However, providing high performance computing environment to the subscribers is very important, so a lightweight anomaly detection method is highly desired.

In this paper, we propose a lightweight IDS with self-similarity measures to resolve these problems. Normally, a regular and periodic self-similarity can be observed in a cloud system's internal activities such as system calls and process status. On the other hand, outliers occur when an anomalous attack happens, and then the system's self-similarity cannot be maintained. So monitoring a system's self-similarity can be used to detect the system's anomalies. We developed a new measure based on cosine similarity and found the optimal time interval for estimating the self-similarity of a given system. As a result, we can detect abnormal activities using only a few resources.

Keywords: cloud computing, information security, self-similarity, lightweight, intrusion detection, anomaly detection.

1 Introduction

Cloud computing is the most innovative Internet-based distributed computing model nowadays. There are four types of cloud computing based on deployment type – public, private, community and hybrid cloud. Table 1 shows the differences between them. Especially, public cloud infrastructure is managed and owned by third party provider. Besides, it has to be accessed through the untrusted and hostile public internet by customers. That is the most obstacles to success of cloud computing era.

To provide secure services for public cloud users, the providers have to deploy security safeguards for auditing and intrusion detection. More than anything else, a

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 353–362, 2011.

Table 1. Cloud computing deployment models [1]

	Infrastructure Managed by	Infrastructure Owned by	Infrastructure Located	Accessible and Consumed by
Public	Third Party Provider	Third Party Provider	Off-Premise	Untrusted
Private/ Community	Organization / Third Party Provider	Organization / Third Party Provider	On-Premise / Off-Premise	Trusted
Hybrid	Both Organization & Third Party Provider	Both Organization & Third Party Provider	Both On-Premise & Off-Premise	Trusted & Untrusted

public cloud system needs to be able to detect internal and external anomalies caused by security failures. However, it is not easy to monitor and detect intrusive events of public cloud systems because there are many virtual or real machines, users and incoming traffics in the cloud. Fig. 1 shows virtual machines (VMs) running in a single hardware. Each VM requires IDS for protecting itself from inside and outside hacker. If a VM is compromised and it generates burst traffic or many processes, then the VM's workload will affect the other VMs on the same real machine. Moreover, some H-IDS in VMs consume much system resources then it will degrade all of the systems in the same real machine. Information security is a key success factor to provide safe cloud computing services. Also, we have consider one of the other key success factor for cloud computing is providing high performance to the subscribers. Therefore, that would be trade-offs between confidentiality versus high performance. That is why lightweight IDS is highly desired in cloud service environment.

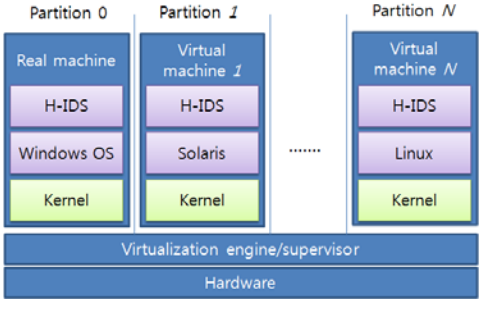


Fig. 1. Architecture of Host based IDS implementation in cloud computing infrastructure

In this paper, we will propose a new self-similarity based anomaly detection method as a lightweight intrusion detection method. We assume that every system has self-similarity during normal status. That means when the system is not under attack, there will be significant trends of all internal events that will have seasonality and repetitiveness. We can observe that every system has its own similarity pattern because the system's internal event patterns depend on the usage pattern of users, processes and various applications. Furthermore, we can observe that outliers on self-similarity values occur when an attack or critical system error happens.

2 Related Work

There are three categories of IDS based on where the sensors are located. While a network-based IDS collects data from network packets, a host-based IDS collects data from the system inside. A hybrid IDS combines a network-based IDS with a host-based IDS. In addition, IDS can be categorized by methodologies that they employ. Misuse detection based IDS works on the signatures of known attacks and thus cannot detect unknown attacks. Anomaly detection-based IDS has the ability to find patterns that do not conform to the expected network or system behavior. So anomaly detection based IDS can respond to unknown attacks.

The IDS we propose uses a method of monitoring self-similarity derived from the cosine distance between event distributions. It has been observed that network traffic and a system's internal activities have self-similarity. The activities can be behaviors by hardware, software and users. If there are some unexpected spikes in activities, they can be seen as related to DDoS attacks, incoming exploits, or malicious user behaviors. So far, researches related to the field of self-similarity are focused on finding anomalies in network protocols such as transfer control protocol (TCP), hypertext transfer protocol (HTTP), and internet protocol (IP), [4, 5] and also in network traffic anomaly detection [3, 6, 7, 8, 9, 10].

We focus on the host's event log analysis because it gives useful information about the system that cannot be captured by network-based IDS. It is better to use a host's event log to create a self-similarity profile because that can reduce useless noise and variables from the system outside, which increases the overall accuracy of the self-similarity profile. In addition, it is the most merit this sensor can work only with the native log files without generating any additional log that can degrade disk I/O performance.

3 Similarity Measure

Because of the needs for a lightweight IDS, we consider which source of information will be most cost-effective. Many host-based IDSs require an audit log created by a basic security module (BSM) for monitoring and tracking system calls and activities. Therefore, the system administrators have to turn on the trusted system convert and run an audit daemon process, but this will increase system CPU load and disk usage drastically. To develop a lightweight IDS, we use the system's default log files in order to avoid deploying an additional logging module. If we detect anomalous events with this default log that do not have plenty of information, that is very easy to apply in cloud computing environment.

3.1 Cosine Similarity Based Self-similarity

3.1.1 Similarity Estimation Process

Suppose g_{xy} is the number of event records when SID_x and $EventID_y$ are given. Then in notation, g_{xy} can be expressed as $g_{xy} = \mathrm{g}(SID_x, EventID_y)$.

Suppose Vector $\overrightarrow{G_t}$ is a system status snapshot in the given time interval *t* (e.g. 30 seconds or 10 seconds) and Vector $\vec{G}$ has elements composed of g_{xy}. So, $\overrightarrow{G_t}$ can be expressed as

$$\overrightarrow{G_t} = (g_{11}, g_{12}, g_{13}, \cdots, g_{1n}, \\ g_{21}, g_{22}, g_{23}, \cdots, g_{2n} \\ \cdots \\ g_{m1}, g_{m2}, g_{m3}, \cdots, g_{mn})$$

where *x*=1,…, m and *y*=1,…, n.

$\vec{E}$ is the same dimensional vector with $\overrightarrow{G_t}$ and it has the values as $\vec{E} = (1, 1, 1, \cdots, 1)$.

Then, we can calculate similarity S_t with cosine distance for each $\overrightarrow{G_t}$ as follows.

$$S_t = \frac{\overrightarrow{G_t} \cdot \vec{E}}{\|\overrightarrow{G_t}\| \|\vec{E}\|}, \text{ where } t=1, \ldots, k. \text{ and } k = \frac{\text{end of time} - \text{begin of time in eventlog}}{\text{giventime interval } t}$$

A is the average value of S_t and σ is standard deviation of S_t.

$$\mathrm{A} = \frac{1}{k}\sum_{t=1}^{k} S_t \,,\ \sigma = \sqrt{\frac{1}{k}\sum_{t=1}^{k}(S_t - A)^2}$$

We define the inequality equation to find the time interval when the outlier occurs.

$$A - 2\sigma \le S_t \le A + 2\sigma$$

We assume that the system's internal activities follow normal distribution when this system is up and running normally. Then, probability an event is within 2 standard deviations of the mean, $P(A - 2\sigma \le S_t \le A + 2\sigma)$, equals 95.4%, so a certain $S_t$out of this range can represent the abnormal event. For all *t*, if S_t satisfies this inequality equation, then the system is normal and has self-similarity. Otherwise, we assume attacks or security violations have happened.

3.1.2 Gaining Optimal Time Interval

To get an accurate result with the method described in section 3.1.1, we need to choose a relevant time interval for estimating self-similarity. That is due to each event having its own seasonality for occurring. Hence, if we cannot choose the optimal time interval adequately, we will find many outliers even though that event has plenty of self-similarity.

To find out an adequate time interval, Let MAXCOUNT be the number of counts that the top number of events generated in the shortest moment. Then the optimal time interval for estimating self-similarity can be obtained as follows.

$$\mathrm{P} = \{\mathrm{p_i} \mid \mathrm{p_i} \text{ is the number of counts bigger than MAXCOUNT / 2}\}.$$

$$\mathrm{T} = \{\mathrm{t_i} \mid \mathrm{t_i} \text{ as the time interval between } \mathrm{p_i} \text{ and } \mathrm{p_{i+1}}\}$$

For all t_i, if t_i satisfies the following inequality, then the optimum time interval is μ.

$$t_i \leq \mu + 2\sigma \ , \ \sigma \leq \frac{\mu}{10}, \ |P| > 3$$

where μ is average of T, σ is standard deviation of T, and $|P|$ is the number of p_i.

3.2 Hybrid Self-similarity Measure

We use the optimal time interval value derived from the method described in 3.1.2. We transform the event log vector $\vec{P}$ to meet the following condition.

$\sum_{i=1}^{n} p_i = 1 \ (p_i \geq 0, i = 1,2,3, \dots, n)$. Then, let $H(\vec{P}) = -\sum_{i=1}^{n} p_i \log_2 p_i$

Now, let $\vec{V}$ is unit vector that has the same dimension with $\vec{P}$

$$\text{where } \sum_{i=1}^{n} v_i = 1 \ (v_i \geq 0, i = 1,2,3, \dots, n).$$

We can estimate similarity value $SIM(\vec{P}, \vec{V})$ as like $SIM(\vec{P}, \vec{V}) = 1 - \beta(\vec{P}, \vec{V})$

$$\text{where, } \beta(\vec{P}, \vec{V}) = H\left(\frac{1}{2}\vec{P} + \frac{1}{2}\vec{V}\right) - \frac{1}{2}[H(P) + H(V)]$$

We can decide that p_i is a normal event, if p_i satisfies the following inequality.

$$\mu - 2\sigma < p_i < \mu + 2\sigma$$

where μ is mean and σ is standard deviation of similarity values. Otherwise, we treat p_i value as outliers. We choose this interval $\mu \pm 2\sigma$ because it is meaningful from statistical view when this distribution follows normal distribution as like S_t.

4 Experiments

We designed the procedures for measuring similarity as can be seen in Fig. 2.

The Windows event log preprocessor extracts the number of events from the Windows' security event log. In the feature selection procedure, our IDS makes groups by combination of SID and EventID. SID is a synonym of Security ID in Windows system. Uniquely, system administrator has 500 value, as like, system guest has 501 value for SID. Table 2 shows the examples of security related event IDs and their meaning. All events related to security are summarized in [11]. In the similarity measurement procedure, our IDS calculates the self-similarity as described in Section 3. If the self-similarity is broken, our system alerts the system administrators. In the check outlier's source procedure, our IDS examines the outlier points and the person who makes the abnormal event or which IP address is the source. Finally, our IDS reports the information to a system administrator and keeps monitoring the self-similarity of the system.

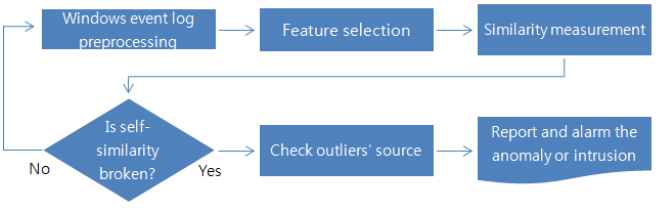


Fig. 2. Procedures for detecting anomalous events with self-similarity measurement

Table 2. Examples of security related event ID

Event ID	Occurrence	Description
529	Logon Failure – Unknown User Name or Password	It indicates an attempt to login with wrong account name or password. When this event repeats a lot, then that can be password-guessing attack by brute-force.
539	Account Locked Out	It indicates an attempt to log on with an account that has been locked out.
627	Change Password Attempt	It indicates that someone other than the account holder attempted to change a password

4.1 Cosine Similarity Function with DARPA the 1999 Dataset

We break down time interval by 30 seconds and remove some noise data. Those noise data are removed because the numbers of records are extremely small, and it does not influence the similarity measurement. Tuesday of the first week does not have any security events at all. The security log file for Friday of the first week has a damaged file so those are ignored as well. Tuesday and Wednesday of the second week have the same log as Monday of that week. Table 3 shows the experiment results with the DARPA 1999 dataset.

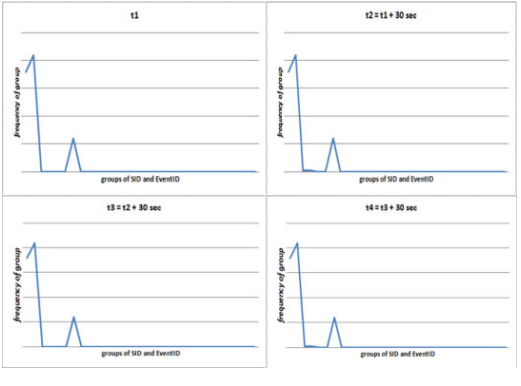


Fig. 3. First 2 minutes of Monday of the first week

Table 3. The experiment results with the DARPA 1999 dataset

Date	A	σ	Outlier ratio
1st Week Mon	0.155244	0.129492	0%
1st Week Tue	No log	No log	No log
1st Week Wed	0.186446	0.173327	0.04%
1st Week Thu	0.167796	0.164791	0%
1st Week Fri	Log error	Log error	Log error
2nd Week Mon	0.348505	0.348505	1.95%
2nd Week Tue	0.348505	0.348505	1.95%
2nd Week Wed	0.348505	0.348505	1.95%
2nd Week Thu	0.088913	0.015611	0.04%
2nd Week Fri	0.244027	0.235173	0%
3rd Week Mon	0.510415	0.015024	0.54%
3rd Week Tue	0.328827	0.027056	0.7%
3rd Week Wed	0.272329	0.267404	0%
3rd Week Thu	0.360024	0.038319	1.09%
3rd Week Fri	0.202278	0.196422	0%

We can find that the DARPA data shows self-similarity overall when an attack has not occurred. Moreover, we can confirm that outliers occur in the second week where there are some attacks. Likewise, we measure the self-similarity of our Windows system's dataset. We simulate a network-based attack with Tenable Nessus Scanner [12]. This vulnerability scanner can emulate attacks selectively and the total number of attack patterns included is 41,014.

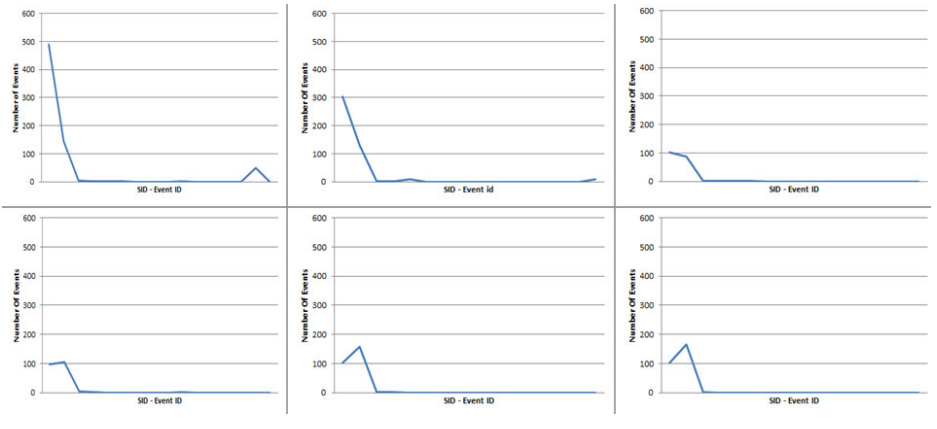


Fig. 4. Self-Similarity changes when the attack occurs

We choose attack plugins within the denial of service, port scanning and windows categories in Tenable Nessus and then it sent 37,484 patterns of attacks. Fig. 4 shows the self-similarity has changed when the attacks are sent to the target system. Also, Tables 4 and 5 show the result of the attack dataset.

Table 4. The result of the attack dataset

Time	S_i	$A \pm 2\sigma$		Situation
t1	0.336748	0.337122	0.370506	Attack
t2	0.343429	0.337122	0.370506	Normal
t3	0.363210	0.337122	0.370506	Normal
t4	0.363706	0.337122	0.370506	Normal
t5	0.350135	0.337122	0.370506	Normal
t6	0.346780	0.337122	0.370506	Normal

Table 5. The result of the attack dataset

Total time	A	σ	Outlier ratio
2 hours	0.353814	0.008346	6.25%

Our system detected the outlier when an attack occurred. Thereafter, we could confirm that self-similarity was re-achieved. In this experiment, the overall false-positive ratio was 4.17 %. When attacks occur, the relevant events are increased, so it is very easy for outlier points to be captured with this algorithm.

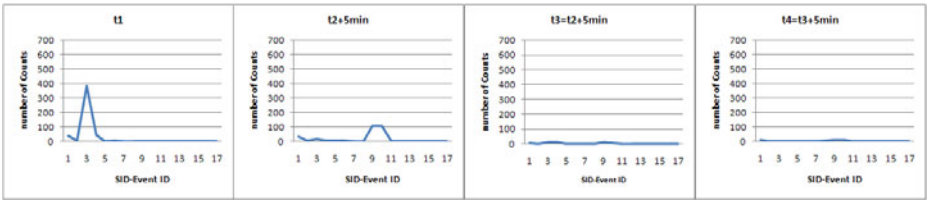


Fig. 5. Self-similarity changes when DDoS attack is incoming

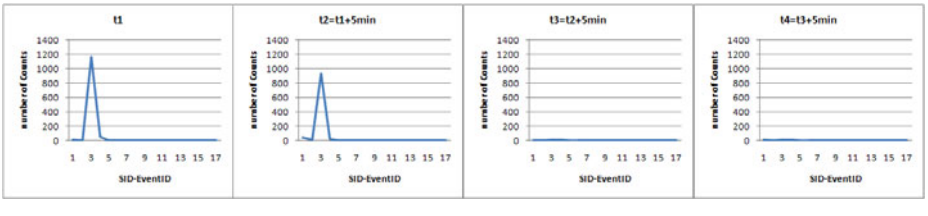


Fig. 6. Self-similarity changes when service-enumerating attack is incoming

Fig. 5 and 6 shows the changes of self-similarity when various attacks are incoming.

We can observe that any attack categories can change self-similarity and we can confirm that with our measure very visually. A false positive can be triggered when some specific service programs are terminated unexpectedly. Even though this is not an attack, it is surely an abnormal event, so it is still useful for an administrator to response such a system malfunction.

4.2 A Statistical Similarity Measure

We only use a cosine similarity based method for estimating self-similarity in 4.1 at the first. We can show better performance with the statistical similarity measures described in 3.2.

Table 6. The experiment result of the DARPA 1999 dataset

Date	A	σ	Outlier ratio
1st Week Mon	0.035786	0.388016	0.1%
1st Week Tue	No log	No log	No log
1st Week Wed	0.139669	0.335384	0.2%
1st Week Thu	0.167796	0.164791	0.1%
1st Week Fri	Log error	Log error	Log error
2nd Week Mon	-0.161480	0.019786	7.5%
2nd Week Tue	-0.161480	0.019786	7.5%
2nd Week Wed	-0.161480	0.019786	7.5%
2nd Week Thu	0.088913	0.015611	0.07%
2nd Week Fri	0.252983	0.238478	0%
3rd Week Mon	0.095879	0.025553	1.0%
3rd Week Tue	-0.194217	0.062902	0.8%
3rd Week Wed	0.307751	0.189072	0%
3rd Week Thu	-0.134530	0.069530	1.1%
3rd Week Fri	0.191734	0.299518	0%

The day without attack, it shows a better outlier detection ability as a result rather than only using a cosine similarity function. It shows much better accuracy in the days when the attack has happened. Table 7 and 8 shows the result of this algorithm when we generate attacks with a Tenable Nessus Scanner.

Table 7. The result of the attack dataset

Time	S_i	$A \pm 2\sigma$		Situation
t1	0.210479	-0.050541	0.111635	Attack
t2	0.167621	-0.050541	0.111635	Normal
t3	0.167621	-0.050541	0.111635	Normal
t4	-0.060631	-0.050541	0.111635	Normal
t5	0.004754	-0.050541	0.111635	Normal
t6	0.044300	-0.050541	0.111635	Normal

Table 8. The result of the attack dataset

Total time	A	σ	Outlier ratio
2 hours	0.030547	0.081088	4.17%

Our system detected the outlier when an attack occurred. Thereafter, we could confirm that self-similarity was recovered again. In this experiment, the overall false-positive ratio was 0 %.

5 Conclusions

The proposed methodology has the following merits to detect anomalous events.

1. It does not need a long learning process that would require many system resources.
2. This self-similarity can be calculated in near-real time.

Our IDS can work robustly even though the Windows event log does not include enough information regarding security rather than the other operating system's audit log. In short, our proposed cosine similarity and hybrid similarity measure is very cost-effective anomaly detection method because it shows high accuracy and it works in near-real time with few system resources.

References

1. Cloud Security Alliance: Security Guidance for Critical Areas of Focus in Cloud Computing v2.1 (2009)
2. McHugh, J.: Intrusion and intrusion detection. International Journal of Information Security 1, 14–35 (2001)
3. Rawat, S., Sastry, C.S.: Network Intrusion Detection Using Wavelet Analysis. In: Das, G., Gulati, V.P. (eds.) CIT 2004. LNCS, vol. 3356, pp. 224–232. Springer, Heidelberg (2004)
4. Crovella, M.E., Bestavros, A.: Self-similarity in World Wide Web traffic: Evidence and possible causes. IEEE/ACM Transactions on networking 5(6), 835–845 (1997)
5. Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V.: Self-similarity through high-variability; statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM Transactions on Networking 5(1), 71–86 (1997)
6. Schleifer, W., Mannle, M.: Online error detection through observation of traffic self-similarity. Proceedings of IEEE Communications 148(1), 38–42 (2001)
7. Allen, W.H., Marin, G.A.: On the self-similarity of synthetic traffic for the evaluation of intrusion detection systems. In: Proceedings Symposium on Applications and the Internet, pp. 242–248 (2003)
8. Li, M., Jia, W., Zhao, W.: Decision analysis of network based intrusion detection systems for denial-of-service attacks. In: Proceedings Conferences on ICII, vol. 5, pp. 1–6 (2001)
9. Nash, D.A., Ragsdale, D.: Simulation of self-similarity in network utilization patterns as a precursor to automated testing of intrusion detection systems. IEEE Transactions on Systems, Man and Cybernetics 31(4), 327–331 (2001)
10. Idris, M.Y., Abdullah, A.H., Maarof, M.A.: Iterative Windows Size Estimation on Self-Similarity Measurement for Network Traffic Anomaly Detection. International Journal of Computing & Information Sciences 2(2) (2004)
11. Microsoft Technet, Security Monitoring and Attack Detection (August 29, 2006), `http://technet.microsoft.com/en-us/library/cc875806.aspx`
12. Tenable Network Security, Nessus, `http://www.nessus.org/nessus/`
13. Wong, S.K.M., Yao, Y.Y.: A statistical similarity measure. In: Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3–12 (1987)

A Production Planning Methodology for Semiconductor Manufacturing Based on Simulation and Marketing Pattern

You Su Mok, Dongsik Park , Chilgee Lee, and Youngshin Han*

School of Information and Communication Engineering SungKyunKwan University 300, Chunchun-dong, jangan-gu, Suwon, Kyunggi-do 440-746, S. Korea
hanys@skku.edu

Abstract. To make production and equipment investments plans in semiconductor lines, the implementation of many variables is needed. These factors bring many changes and the result is hard to predict. It is difficult to define a standard because there are many influencing factors to make a prediction. This project established semiconductor production plans using the marketing pattern references on the past to satisfy all conditions from the factors. We come up with thesis on reasonable and standardized processes.

1 Introduction

Semiconductor industry is the core-part industry leading the society to information community of the 21st century. This will accelerate to the latest industry development of the nation. Timing industry, higher additional value, higher growth, higher risk, and technical integration are the distinctive feature of the semiconductor. Therefore, precise prediction of the market is an important factor to respond with most suitable production plans for this situation[1][9][12]. But there variety of restrictive conditions, it is difficult to satisfy these conditions with a production plan. Until now, people spent too much time on these tasks, therefore automation processes is needed. We suggest a reasonable method of producing semiconductor using pattern data from the past to solve this problem.

This paper is organized as following. Chapter 2 describes a plan on production of semiconductor. In Chapter 3, presents planning line production using pattern data. Chapter 4 concludes with directions for further research.

2 Production Plan for Semiconductor

Establishment of sales plan

"Establishment of sales plan" is about how much products will be sold in a certain point. The sales plan relies on the following [4][6][7][11]: demand of products, income of the company, possession of the market share and sales cost of each product.

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 363–371, 2011.

Design Rule (DR)

Process ability and method of manufacturing should be considered in the process of designing semi-conductor. Each functional part and the physical distance covers the contents of regulations of design. Table 1. shows design rule. The product can be produced using many different DR, and applying DR is different with each product. The number of chips which can be made from a wafer differs with each DR.

Table 1. Amount of chips produced

Product	DR	Amount of chips produced
A	DR1	100
	DR2	120
	DR3	150
B	DR2	200
	DR3	220

Table 2. Estimating the producible capacity by line

Line	DR	Capacity
1 Line	DR1	20000
	DR2	25000
2 Line	DR2	15000
	DR3	30000

Release period by products and by DR

DR is different for each product, and the time for the DR to be applied to the line is different. Therefore, the production date by product and by DR is needed to establish the plan.

Estimating the producible capacity by line

Each line has a producible capacity but the usable DR differs from the equipment in possession. Therefore to estimate the producible capacity, capacity should be estimated by line and by DR. An example is shown in Table 2.

Development plan by product, by DR

To figure out the development period by product and by DR, understanding the period to use the line is needed.

Estimating priority of the production

Priority of the production is the ratio of DR by product. For example, in A product, there are DR1, DR2, and DR3 which production can be applied to. Then priority of the production can be estimated as shown in Table 3 and the sum of all DR should be 100%.

Table 3. Estimating priority of the production

Product	DR	Priority of products (%)
A	DR1	30
	DR2	20
	DR3	50
	Sum	100

Estimating priority of the product is the intermediate result before reaching the final result. To get a precise result from the line distribution, priority of the product estimation should be exact. There are many algorithms to get the result for priority of the product. In this paper, pattern base algorithm is used to get the result of the line distribution using priority of the product.

Line Distribution

All elements produced above should be distributed appropriately to the DR of the line concerning the producible capacity. If the elements are distributed unconditionally to the DR even if it has enough capacity, many problems can be generated. We assume that there are products produced as show in Table 4 in the first quarter of 2005.

Table 4. Necessity amount of the wafer

Product	DR	Amount needed to be produced
A	DR1	20000
	DR2	10000
	DR3	10000
B	DR2	15000
	DR3	20000

We assume that the producible capacity by line and by DR is as described in Table 5.

Table 5. Capacity by line and by DR

Line	DR	Capacity
1 Line	DR1	20000
	DR2	25000
2 Line	DR2	15000
	DR3	30000

If the distribution behaved like explained in Table 6., the production and necessity amount of the wafer can be produced within the producible capacity.

Table 6. Capacity by line and by product

Line	Product	DR	Capacity
1 Line	A	DR1	20000
	A	DR2	5000
	B	DR2	5000
2 Line	A	DR2	5000
	B	DR2	10000
	A	DR3	10000
	B	DR3	20000

The production line management is planned by a quarter of a year, and the operation above is done for every quarter. The described above operation has several problems.

First, there are 10s and 100s of lines, products, and DRs. For every element many other elements related. If one is changed many elements related to the changed elements have to be changed.

Second, the amount of products produced in line by product and by DR should be continuous through out the quarters. Each product is not only produced. The conditions of the equipments in line should be considered. If the amount of production needed to be increased or decreased, the equipments in a semiconductor line should be upgraded or replaced. If these conditions are ignored with unreasonable increase or decrease, problems will occur. For example, there was a plan to produce 500 products this quarter. In the following quarter, the plan should be within +800 ~ -200 products for the process to continue without many problems.

Even if the required amount of each product that can be produced within the producible capacity of the line, may not be the same in reality. For example, if the producible capacity of DR1 in a whole line is 20000, and the amount of products needed is only 10000, and if the producible capacity of DR2 is 35000, and the amount of products needed is 50000. In these cases, there will be problems. For these cases, the amount of products can be changed, and if there aren't enough resources, new line should be built.

With these factors, planning is a very hard task even for experts. From this research, theory of best suitable plan is made. This theory satisfies many of these restricted conditions using data patterns line distribution in the past.

3 Line Distribution Using Pattern Data

3.1 Definition of Pattern

Pattern of DR and products can be analyzed with production data in the past. Generally, there are specific difference between patterns of general products and special products. The reason is that general products follows cycle of the common market, but the special products are produced and sold by the result of businesses.

In the beginning of the graph, the amount of the product increased enormously increase. The graph maintains the rate for certain amount of time and it slowly decreases.

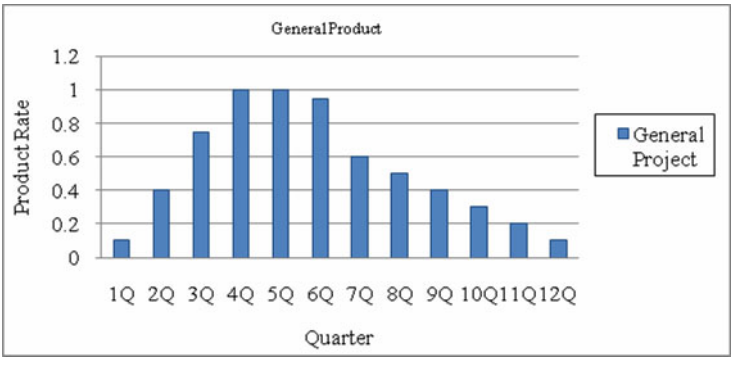


Fig. 1. Production data pattern in general products

Generally, standardized general products are produced for 3 years. But the period of the production can be reduced. Therefore, this paper uses data patterns of general products of 2 years and 2.5 years.

3.2 Release Period and the Amount of Production from Prior Quarter

Pattern is a relative amount of a product, not an absolute amount. Therefore the amount of production in a specific quarter is references the amount of production from prior quarter, release period of the product, and the patterns of the past.

3.3 Distribution of the Line

From the above data, many elements are combined and distributed to the line. Example is as follows. Let's assume that the amount of production by line, by products, and by DR in the first quarter of 2004 is as in Table 7.

Table 7. Distribution of the line

Line	Product	DR	First quarter of 2004
1 Line	A	DR1	1000
	A	DR2	500
	B	DR2	1000

The period of release is as in Table 8.

Table 8. Period of release

Line	Product	DR	Period of release
1 Line	A	DR1	2003year 1 quarter
	A	DR2	2003year 4 quarter
	B	DR2	2003year 3 quarter

If the products produced above are assumed as general products, then the data pattern of general product production should be applied as in Table 9.

Table 9. The data pattern of general product production

1Q	2 Q	3 Q	4 Q	5 Q	6 Q	7 Q	8 Q	9 Q	10 Q	11 Q	12 Q
0.1	0.4	0.75	1.0	1.0	0.95	0.6	0.5	0.4	0.3	0.2	0.1

Amount of products in first quarter of A product DR1 in 1 line is 1000, period of release is the first quarter of 2003. If the pattern data is applied, it is equivalent to 5Q 1.0. Value of 6Q in next quarter is 1.0, same as 5Q, therefore the amount of production of A product DR1 from 1 line is 1000. The amount of production of A product DR2 in the first quarter is 500, and the period of release is fourth quarter in 2003. If

above data is applied to the pattern data, then the data corresponds to 0.4 of 2Q. The value of 3Q of next quarter is 0.75. Therefore, the amount of second quarter = 500 * 0.75 / 0.4 which equals to 937. If the rest of B product is calculated in the same manner, then the amount of products will look like Table 10.

Table 10. The amount of quarter after distribution of the line

Line	Product	DR	First quarter	Second quarter
1 Line	A	DR1	1000	1000
	A	DR2	500	937
	B	DR2	1000	1333

3.4 First Redistribution

The distribution process of line above, capacity has not been considered. Distribution of line has been applied using pattern data and the amount of previous products. In this case, amount of wafer produced in line can be over or under the limitation.

But the production of the wafer in line is efficient to use 100% of the capacity. Discordance should be resolved by redistribution. Let look at another example with redistribution of line.

Table 11. The amount of quarter

Line	Product	DR	First quarter	Second quarter
1 Line	A	DR1	1000	1000
	A	DR2	500	937
	B	DR2	1000	1333

Table 12. The Capacity of quarter

Line	DR	First quarter of capacity	Second quarter of capacity
1 Line	DR1	1000	1000
	DR2	1500	1600

As in two Table (11 & 12) above, DR of line can be produced by the capacity in the first quarter. But if the pattern data is applied to amount of production in first quarter to calculate the amount of production of second quarter, it is within the capacity of DR1, but the capacity of DR2 is over by 670 sheets. In this case, capacity of DR in line should be redistributed to produce properly.

The redistribution algorithm is as follows.

The amount of exceed capacity of DR and products should be calculated. Distribute the exceeded amount of wafers to corresponding DR with enough capacity.

Let's look at another example concerning other lines.

Table 13. Production result

Line	DR	Second quarter of capacity	Second quarter of production	Result
1 Line	DR1	1600	2270	+670
2 Line	DR2	1200	1000	-200
3 Line	DR3	1670	1200	-470

In Table 13, capacity of the DR2 in 1 line is 1600. 2270 should be produced, therefore this equipment cannot produce 670. But the capacity of DR2 in 2 line is 1200 and it only produces 1000, so it can produce 200 more products. 3 line can product 470 more products. Therefore, the exceeding requirement of production in 1 line should be produced in 2 line and 3 line.

Second Redistribution

By using the first redistribution algorithm, solved the conflict between the capacity and the amount of products. But it cannot always satisfy the situation. In this case, the next algorithm is used for redistribution.

Even if the line uses different DR, if it produces same product, then distribute wafers to that line. If DR is different, difference of NetDie should be considered.

Distribution Ratio

We have looked at the first and the second redistribution. Redistribution is applied to perform the maximum operation of the line even if it ignores the pattern. Producing products have discontinuity by line, by product, nor by DR. Discontinuity could be applied to some degree, but if it is applied excessively, then this method would be reject from the field. To maintain the flow of the process to some degree, Redistribution Ratio is applied.

Redistribution Ratio

The limitation ratio of moveable product to prevent sudden increase or decrease of certain products in specific DR of specific line.

If the Redistribution Ratio is applied, then the continuity is increased but the utilization of line capacity can decrease. The utilization is applied in producing each product to some degree. So in this result of this experiment, by applying Redistribution Ratio, more practical distribution plan could be organized.

3.5 Experiment Result

Amount of products by line, by products, and by DR is as follows. The y-axis shows the amount of production produced in a line. Increase in amount of products produced is shown in the graph. The production period is specific by products, by DR, and by amount of production. The result of the model is compared to the result of actual production plan. The difference comes from the elaborateness of the pattern, but overall, the result shows more than 96% of similarity.

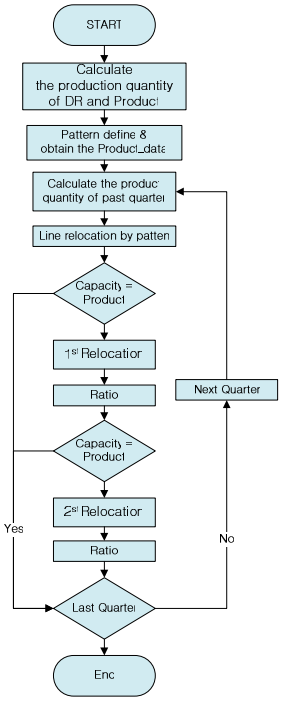


Fig. 2. Flow chart of standard process

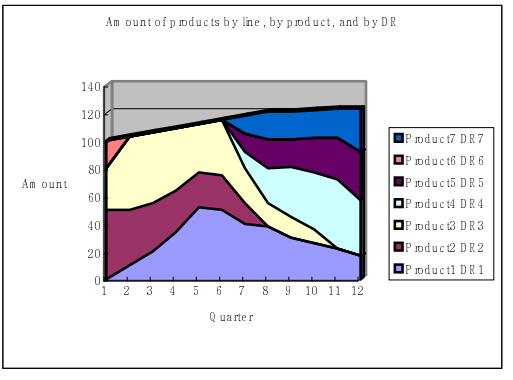


Fig. 3. Amount of products by line, by products, and by DR

4 Conclusion

This experiment has used the pattern made from the market data of the past to plan investment plan of equipment and to plan production of line. There were some problems which occurred during the experiment. To apply to pattern data from the past to plan production in the future was difficult. The pattern which should be applied to the products has been decided in subjective manner. It is unreasonable to apply just the pattern data from the past because the period of production is decreasing. Therefore the pattern is not just used. The result has been extracted from different period. The computing process takes only 3 seconds, but it took much time to adjust the pattern. If this method is compared with the method used in the past, it is certain that the time is decreased from 10 days to 3 hours. By using the left over time to test in many other methods, more efficient plan of operation the line has been made.

Amount of productions has been calculated in this paper using producible capacity of line by quarters which is already decided. The producible capacity of line should consider the increase of equipments and line, but it wasn't considered in this paper. In the future, experiment of the increase of equipments applied to line production plan should be performed.

Acknowledgments

This work is supported by Basic Research Program through the National Research Foundation of Korea(NRF) funded by Ministry of Education, Science and Technology(2010-0003149).

References

[1] Choi, B.K., Kim, B.H.: MES (manufacturing execution system) architecture for FMS compatible to ERP(enterprise planning system). INT. J. Computer Integrated Manufacturing 15(3) (2002)
[2] Lee, Y.H., Kim, T.: Manufacturing cycle time reduction using balance control in the semiconductor fabrication line. Production Planning and Control 13(6), 529–540 (2002)
[3] Lee, Y.H., Park, J., Kim, S.: Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. IIE Transaction 34, 179–190 (2002)
[4] Park, D., Han, Y., Lee, C.: Optimization of a simulation for 300mm FAB semiconductor manufacturing. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, pp. 260–268. Springer, Heidelberg (2006)
[5] Han, Y., Lee, C.: RRAM spare allocation in semiconductor manufacturing for yield improvement. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 95–102. Springer, Heidelberg (2004)
[6] Han, Y., Park, D., Chae, S., Lee, C.: Full fabrication simulation of 300mm wafer focused on AMHS (Automated material handling systems). In: Baik, D.-K. (ed.) AsiaSim 2004. LNCS (LNAI), vol. 3398, pp. 514–520. Springer, Heidelberg (2005)
[7] Potoradi, J., Boon, O.S., Mason, S.J.: Using simulation-based scheduling to maximize demand fulfillment in a semiconductor assembly facility. In: Proceedings of the 2002 Winter Simulation Conference, pp. 1857–1861 (2002)
[8] Rupp, T.M., Ristic, M.: Fine planning for supply chains in semiconductor manufacture. Journal of Material Processing Technology 107, 390–397 (2000)
[9] Vargas-Villamil, F.D., Rivera, D.E.: Multilayer optimization and scheduling using model predictive control: application to reentrant semiconductor manufacturing lines. Computers and Chemical Engineering 24, 2009–2021 (2000)
[10] Vargas-Villamil, F.D., Rivera, D.E., Kempf, K.G.: A hierarchical approach to production control of reentrant semiconductor manufacturing lines. IEEE Transactions on Control systems Technology 11(4), 578–587 (2003)
[11] Kim, S., Yea, S.H., Kim, B.K.: Shift scheduling for steppers in the semiconductor wafer fabrication process. IIE Transactions 34, 167–177 (2002)
[12] Hsieh, B.W., Chen, C.H., Chang, S.C.: Scheduling semiconductor wafer fabrication by using ordinal optimization-based simulation. IEEE Transactions on Robotics and Automation 17(5), 599–608 (2001)

Data Hiding in a Halftone Image Using Hamming Code (15, 11)

Cheonshik Kim, Dongkyoo Shin, and Dongil Shin

Department of Computer Engineering, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul 143-747, Korea
mipsan@paran.com, shindk@sejong.ac.kr, dshin@sejong.ac.kr

Abstract. This paper presents a data hiding technique for a halftone image. Each block of halftone bitmaps is transformed into a sequence of binary bits, and then regarded as a codeword. From the codeword, we can get a syndrome value. The XOR operation between a syndrome and 4-bits secret message is used to conceal 4-bits in the codeword. If the value of a XOR operation is not decimal zero, the position in a codeword should be flipped. In this way, one can hide a message. When every embedding procedure is finished, a sender transmits a stego halftone image to a receiver. A receiver can then extract secret data with a stego image, using a hamming code scheme. Using this procedure, we can conceal secret data in a halftone image and vice versa. Our proposed method is to hide 4-bits in a 4×4 block to flip a bit, is easy to implement, and achieves a high embedding capacity with good perceptual quality. It can be used to copyright and secret communications.

Keywords: BTC, Error Diffusion, Dithering, Hamming Code, Halftone Images.

1 Introduction

Recently, data hiding has become a very interesting field for security and copyright, which embeds information into an image. The embedded data usually contains ownership identification, authentication data, and other information useful for different applications such as copyright shield, data integrity authentication, verification of origin of data, recipient tracking, and more. The main requirement for various applications is that the embedding change the image content imperceptibly. The halftone [7] technique can be used to print books, newspapers, and magazines because such printing does not require as good a quality as natural color. A halftone is used to produce two-tone texture patterns as approximations of the original multi-tone images [1-6].

There are two main kinds of halftoning techniques, ordered dithering [1, 6] and error diffusion [2, 5, 6]. Error diffusion is more complicated than ordered dithering, but it can yield a higher visual quality with few blocking effects. It compares the sum of image pixel intensity and error from the past with a mixed threshold to determine the output. The halftoning error is then fed forward

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 372–381, 2011.

to its adjacent neighbors using a kernel, so each image pixel efficiently has an adaptable threshold.

Two commonly used kernels are the Jarvis and the Steinberg [5]. The Jarvis kernel has large support and tends to give halftone images with high contrast and coarse texture. The Steinberg kernel has smaller support and gives halftone images with fine texture and good contrast. Although a few bit-planes are changed in a 4×4 block in halftone image, it will produce a bad effect in the quality of an image. Thus, it is important to flip a bit in a 4×4 block of an image, if possible.

However, most of the scheme is not proper for hiding much data in a halftone image because a halftone image is regarded as a binary image. Bit planes are composed of 0's or 1's. Therefore, the special techniques suitable for halftone images use covering code [10, 11, 12], hamming distance [1] and modified EMD methods [14]. The reference, [10, 11, 12] used hamming code for data hiding. Zang [12] and Chang [13] applied their algorithms to grayscale images. On the other hand, Chao[11] proposed a method to introduce data hiding into a halftone image. Chao's method is used on a halftone image and a hamming code (7, 4) to hide a secret data. This method provides reasonable capacity, but the image quality is bad for actual application.

In this paper, we present an improving scheme of hamming code for data hiding in a halftone image, so it can enhance both capacity and quality. The paper is organized as follows: In Section 2, Error Diffusion algorithms are briefly described. The proposed hamming code scheme is illustrated via examples and algorithms in Section 3. In Section 4, experimental results are offered. Finally, conclusions are discussed in Section 5.

2 Related Works

In this chapter, we explain Error-Diffusion algorithm related data hiding in a halftone compressed image. Error-Diffusion is a dithering algorithm and is used to publish newspapers, magazines, and books.

2.1 Error Diffusion

Error diffusion generates blue noise, which is a high frequency complement of pink noise. It produces good quality halftone patterns, but low frequency energy is the enemy. Therefore, the blue noise concept has become an important part of halftoning research [6].

In Fig.1 (a), assumes that the input signal varies from g=0.0 (black) to g=1.0 (white). The threshold is 1/2, so block simple sets the output to 0 for values less than 1/2 and to 1 for values greater than or equal to 1/2. The binary output signal is subtracted from the pre-threshold signal to form an error. This error is "diffused" into yet to be considered input values as governed by the Error Filter. This algorithm was introduced by Floyd and Steinberg [5], who also proposed the error filter shown in Fig. 1(b).

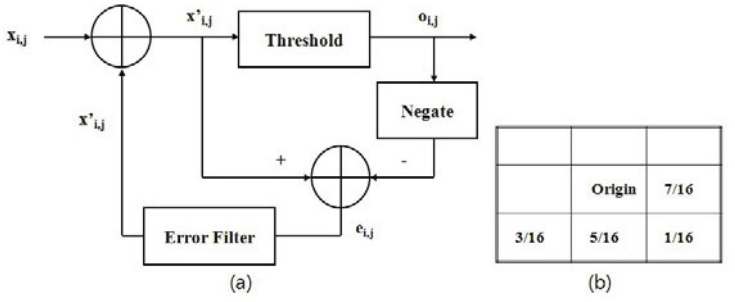


Fig. 1. (a) The error diffusion algorithm. (b) Floyd and Steinberg (Kernel)

2.2 Modified Error Diffusion [9]

The original grayscale is divided into $n \times n$ blocks. In the original BTC method, the image was segmented into 4×4 pixels blocks, assuming that $x_1, x_2, \ldots, x_{n\times n}$ are values of the pixels in a block. For each block, the mean value ($\overline{x}$), maximum (x_{max}), and minimum (x_{min}) are then calculated and encoded.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n^2} x_i \tag{1}$$

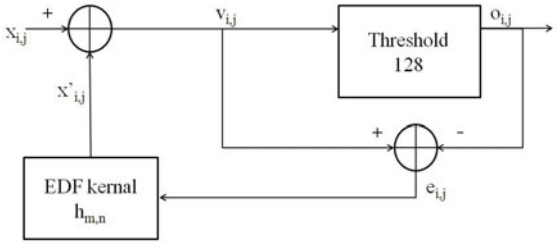


Fig. 2. shows the diagram for the modified error diffusion scheme

The variable $x_{i,j}$ denotes the current input pixel value, and $\overline{x}_{i,j}$ denotes the diffused error sum added from the neighboring processed pixels with the kernel. The variable $o_{i,j}$ denotes the binary output in position (i, j), and the error kernel $h_{m,n}$ is used to diffuse the error caused by the difference, $e_{i,j}$, between the output binary value and the input grey level value. The kernel is shown in Fig. 1(b), where the origin denotes the position of the currently processed pixel. The variable $v_{i,j}$ denotes the modified value. The variables can be calculated as follows by Eq.(2) and Eq.(3).

$$v_{i,j} = x_{i,j} + \overline{x}_{i,j}, \quad where \quad \overline{x}_{i,j} = \sum_{m=0}^{2} \sum_{n=-2}^{2} e_{i+m,j+n} \times h_{m,n} \tag{2}$$

$$e_{i,j} = v_{i,j} + o_{i,j}, \quad where \quad o_{i,j} = \begin{cases} x_{min} & if \quad v_{i,j} < \overline{x} \\ x_{max} & if \quad v_{i,j} \geq \overline{x} \end{cases} \tag{3}$$

In the Error Diffusion method, high mean and low mean is selected for decoding, so it is a time complex method. However, Modified Error Diffusion employs the maximum and minimum values of the block. Thus, this method reduces time complexity and improves image quality significantly.

3 Proposed Method

In this section, we introduce the hamming code theory and propose a data hiding method using hamming code. A halftone image is the size of $N \times N$, which is composed of stream of n-pixels, i.e., $\{0, 1\} \in n$, because a halftone is a bitmap image. Let r be a non-negative integer, the dimension of the parity space.

Let $n = 2^{r-1}$ be the code length and $k = n - r$ be the number of bits that is encoded in each codeword [11, 12, 13]. The codeword will have a minimum Hamming distance of $d = 3$, so that one error can be corrected, two errors detected. Note that one can argue that in order to correct one error, the errors bit position must be determined. For an n bit code, $log_2 n$ bits are, therefore, required. The parity check matrix for the [7, 4] Hamming code is Eq.(4).

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{4}$$

For c to be a codeword, it must be in the null space of this matrix, i.e., $\mathrm{H}c = 0$. We are assuming that there is a sequence of bits, which has an error in the first bit position, i.e., 1101010_b. We calculate the syndrome as,

$$S = H \times (c)^t \tag{5}$$

where, H is the parity checker matrix and c is a 7-bits sequence binary number. That is, the syndrome is $([011])^t$. If a syndrome value is non-zero, it denotes the position of the bit error. If you are flipping a bit of the position value in a codeword, the codeword will be the correct bits. Binary Hamming codes are $[2^r-1, 2^r-1-r]$ linear codes with parity check matrix H of dimensions $r \times (2^r-1)$ whose columns are binary expansions of numbers $1, \ldots, 2^r - 1$. For example, the parity check matrix H for r=4 is

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{6}$$

Let us assume that the cover object is an image consisting of $N \times N$ pixels. To use matrix embedding, divide the cover image into $N \times N$ blocks with non-overlap and scan from left to right and from top to bottom, each pixels consisting of n, where n is block size.

Example 1. We assume that message m=([101]) and codeword c=([1101001]). It is easy to calculate the syndrome using Eq.(5) with parity checker H and a codeword. The syndrome is $H \times (c)^t = ([000])^t$. In order to hide a message in a codeword, an exclusive operation is computed in between vector $(c)^t$ and message m, i.e., $w = (H \times (c)^t) \oplus m$. Next, if the computed syndrome vector w is 0, there is no need to flip a pixel. Otherwise, find the w-th column of c and flip the w-th pixel. In this example, since the w is not zero, the 5-th ([101]) position must be flipped.

3.1 Embedding Algorithm

For secret communication, a halftone image must conceal a message that a sender wants to transmit. In this section, we will show the embedding algorithm stage by stage.

Input: Original halftone image HI with $N \times N$ pixels and the secret data δ, parity check matrix H. **Output:** A stego image SI, the peak point α, the minimum point β, length of secret data $|\delta|$.

Step 1. First, we construct a halftone image, using Eq.(2) and (3) with a grayscale image. $N \times N$ is a block size, which is a unit of processing halftone. $x_{i,j}$ is a pixel value in a 4×4 block of grayscale image. x_{max} is the largest value in a block, and x_{min} is the smallest value in a block. $o_{i,j}$ takes the value of $v_{i,j}$ depending on Eq.(3). If a $v_{i,j}$ is greater than the $\overline{x}$, a $o_{i,j}$ is assigned to '1'; otherwise a $o_{i,j}$ is assigned to '0'. When a block is encoded to a halftone image, x_{min} is '0' and x_{max} is '1'. Thus, we get a halftone block in Fig.3. Divide the halftone image HI into 4×4 blocks, which are composed of a $c = \{c_1, c_2, \cdots, c_{16}\}$, i.e., $c_i \in \{0, 1\}$, where $i = \{1 \leq i \leq 16\}$.

Step 2. Read a block and calculate the syndrome using Eq.(5) with parity checker H and c, i.e., syndrome is $S_1 = H \times (c)^t$, where $c = \{c_1, c_2, \cdots, c_{15}\}$.

Step 3. Read a n-bit binary data; b_n is extracted bits from the secret data δ, where n=4 and $b = \{b_1, b_2, b_3, b_4\}$, where $b_i \in \{0, 1\}$.

Step 4. Compute the exclusive operation (XOR) between 4-bit secret data and syndrome value from a block, i.e., $S_2 = \text{XOR}(b_i, S_1)$.

Step 5. A syndrome, S_2 denotes an index of a codeword, i.e., index is a position, which takes place as an error. If S_2 is zero, it means that there is no error in

a codeword. Otherwise, we should flip the value of position in a codeword. Through these steps, it can be possible to hide 4-bits of binary secret data.

Step 6. Go to Step 2 to continue the embedding processes, until there exists no message for embedding in a halftone image.

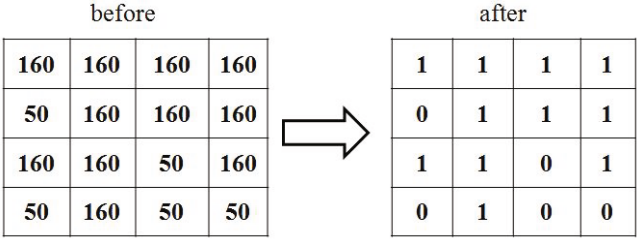


Fig. 3. (a) a block of halftone, (b) a bit pane

Example 2: Codeword $c = [1\ 0\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 1]$, which is the bits of Fig. 3(b), reading from top to bottom and from left to right. A message is $b_i = [1\ 0\ 0\ 1]$. Calculate $S_1 = H \times (c)^t = [0\ 0\ 0\ 1]$ and $S_2 = \mathrm{XOR}(b_i, S_1) = [1\ 0\ 0\ 0]$. The S_1 and S_2 is the syndrome value, and S_1 is not including message value. The S_2 is the applied XOR operator between S_1 and message b_i. S_2 indicates we should flip a bit in a codeword, when a syndrome is not zero. In this example, the first position in a codeword should be flipped as in Fig.4. As a result, a receiver can possbily extract a message from the stego image, by calculating the syndrome of the hamming code.

In a 4×4 block, there is only 1-bit flipped when embedding 4-bits. For hiding a message in a codeword, we used a halftone image composed of bit-planes, i.e., $\{0, 1\}$.

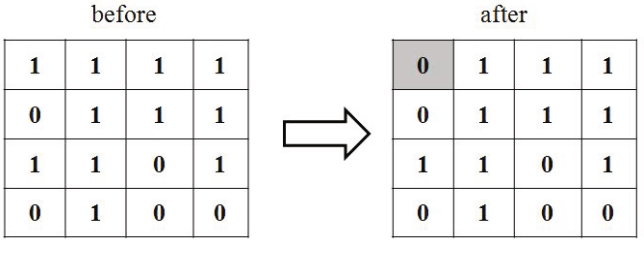


Fig. 4. An example of hamming code in an image

3.2 Extraction Algorithm

After the embedding procedure is complete, a sender transmits a stego image with secret information to a receiver through the communication channel. A receiver will find a private message with a stego image, using the extracting algorithm.

The extraction procedures are below:
Input: The stego image SI with $M \times N$ pixels and parity check matrix H.
Output: The original secret message δ.

Step 1. Divide a halftone image SI into 4×4 blocks, composed of a $c = \{c_1, c_2, \cdots, c_{16}\}$, i.e., $c_i \in \{0, 1\}$, where $i = \{1 \leq i \leq 16\}$.
Step 2. $cnt = 0$, count number of blocks, i.e., $bit_c = cnt \times 4$, where, bit_c is numbered of bits to be concealed in an image.
Step 3. Read a block and calculate the syndrome using Eq.(5) with parity checker H and c, i.e., syndrome is $S = H \times (c)^t$, where $c = \{c_1, c_2, \cdots, c_{15}\}$.
Step 4. The syndrome S is a 4-bit piece of secret data. It can be combined as $\delta = \delta||(S)$. $cnt = cnt + 1$.
Step 5. Repeat step 3 until $cnt = |\delta|$.

4 Experimental Results

We had experimented with data hiding using nine 512×512 halftone images, which were obtained by using the Error Diffusion Dithering algorithm [2] from 8-bit gray level images. The important factor for stego image is the quality and capacity of an image. A PSNR is required to measure a stego image quality against its original image. The quality of a stego image is an important factor to apply for communication and watermarking.

A PSNR assumes that distortion is only caused by additive signal-independent noise. Therefore, noise measures applied directly to a restored image and its original do not gauge the visual quality. Quality measures based on linear HVS models assess image quality in three steps. First, an error image is calculated as the difference between the original image and the restored image. Second, the error image is weighted by a frequency response of the HVS given by a low pass contrast. Finally, a signal-to-noise ratio is computed. These quality measures can take into account the effects of image dimensions, viewing distance, printing resolution, and ambient illumination. They do not include non-linear property of contrast perception, such as local luminance, contrast masking and texture masking. A PSNR was developed to measure the quality of a grayscale image. Thus, the PSNR was not used to measure our scheme for a halftone image directly. For this reason, Guo modified the PSNR equation as Eq.(7). Let the size of the original image be P×Q. The error criterion involved in this study is defined as below.

$$PSNR = 10log_{10} = \frac{(P \times Q \times 255^2)}{\sum_i^P \sum_j^Q [\sum_{m,n \in R} \sum w_{m,n}^2 (x_{i+m,j+n} - \varepsilon_{i+m,j+n})^2]}, \quad (7)$$

Fig.5 shows the comparison of image quality between a proposed scheme and previous schemes. As shown in Fig.5, our scheme provides much higher image quality than do other schemes. Tsai [10] used the VQ method to make a halftone image. Therefore, the key problem for the noise introduced by compression is to reproduce the details removed by down-sampling. The VQ method does not

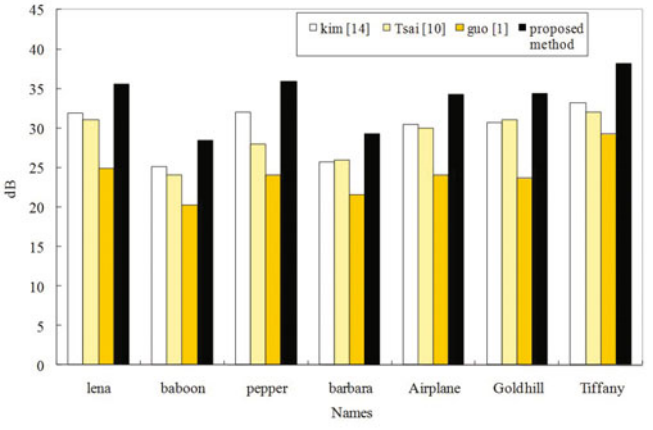


Fig. 5. Comparison of image quality for previous and proposed schemes

provide high quality image. Kim [14] also used IODBTC (Improved Ordered BTC) halftone and concealed secret data using the EMD algorithm [14]. They show good image quality because only two pixels in a 4×4 block are flipped. With the proposed scheme, it can be possible to flip only one pixel in a 4×4 block. Therefore, our scheme is better than that for any other method.

Table 1. shows that the embedding capacity of the proposed scheme is larger than the previous schemes, such as PAN[2], Tsi[10] and Kim[14]. PAN[2] used the LUT (Look-Up Table) for data hiding.

The Look-up table based method [2] requires cost, which is the construction and search of LUT with an image. Thus, it was a time complex method. Tsi [10] proposed reversible data hiding for vector quantization a compressed image, which required a codebook that reduced the size of an image. This scheme applied the histogram modification method to a codebook. Our proposed scheme used hamming code (15, 11), where it is possible to hide 4-bit secret data as a 1-pixel flip in a 4×4 block.

In Fig.6, we can see that stego images can be produced using the proposed scheme. As you can see, it is not easy to discriminate between a halftone and a

Table 1. Comparison of embedding capacity between previous and our method

	Capacity			
Images	PAN[2]	Kim[14]	Tsi[10]	Proposed Method
Lena	831	42,129	4,236	65,536
Airplane	1191	42,088	5,249	65,536
Baboon	54	42,136	698	65,536
Boat	553	42,192	6,305	65,536
Pepper	685	42,127	5,037	65,536
Barbara	254	42,166	3982	65,536

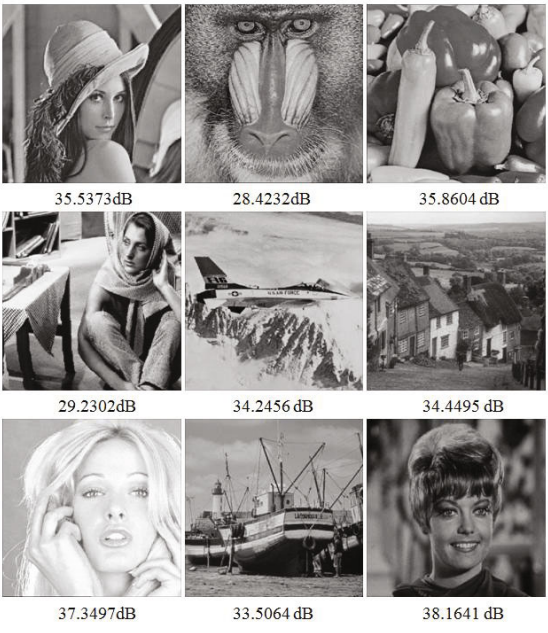


Fig. 6. The quality of images in the proposed schemes

stego image. The experiment result shows that our proposed scheme is a novel method for data hiding in a halftone image. Moreover, it is possible to use steganography, because a stego image will have as good quality as a grayscale image.

5 Conclusion

In this paper, we proposed a data hiding scheme for a halftone image, using hamming code (15, 11). The proposed scheme has good performance against embedding capacity and the quality of a halftone image. Chao[11]'s method is hamming code(7,4) for data hiding, which produces bad quality of a halftone image. On the other hand, our scheme show good quality of a halftone image, thus appropriate for various fields, such as copyright, personal protection, military, and communications, because we reduced the distortion.

Acknowledgement

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2009.

References

1. Guo, J.M.: Watermarking in dithered halftone images with embeddable cells selection and inverse halftoning. Signal Processing 88, 1496–1510 (2008)
2. Pan, J.S., Luo, H., Lu, Z.H.: Look-up Table Based Reversible Data Hiding for Error Diffused Halftone Images. INFORMATICA 18(4), 615–628 (2007)
3. Tu, S.F., Hsu, C.S.: A BTC-based watermarking scheme for digital images. Information & security 15(2), 216–228 (2004)
4. Tseng, H.W., Chang, C.C.: Hiding data in halftone images. INFORMATICA 16(3), 419–430 (2005)
5. Floyd, R.W., Steinberg, L.: An adaptive algorithm for spatial grey scale. In: Proceedings of the Society of Information Display, vol. 17, pp. 75–77 (1976)
6. Ulichney, R.: A Review of Halftoning Techniques, Color Imaging: Device-Independent Color, Color Hardcopy, and Graphic Arts V. In: Proc. SPIE, vol. 3963 (January 2000)
7. Delp, E., Mitchell, O.: Image Compression Using Block Truncation Coding. IEEE Transactions Communications 27, 1335–1342 (1979)
8. Niranjan, D.V., Thomas, D.K., Wilson, S.G., Brian, L.E., Alan, C.B.: Image Quality Assessment Based on a Degradation Model. IEEE Transactions on Image Processing 9(4) (2000)
9. Guo, J.M.: Improved block truncation coding using modified error diffusion. Electronics Letters 44, 462–464 (2008)
10. Tsai, P.: Histogram-based reversible data hiding for vector quantization-compressed images. Image Processing, IET 3, 100–114 (2009)
11. Chao, R.M., Ho, Y.A., Chu, Y.P.: Data Hiding Scheme Using Covering Codes in Halftone Images Based on Error Diffusion. In: Asia-Pacific Services Computing Conference, APSCC 2008, pp. 1483–1488 (2008)
12. Zhang, W., Wang, S., Zhang, X.: Improving Embedding Efficiency of Covering Codes for Applications in Steganography. IEEE Communications Letters 11, 680–682 (2007)
13. Chang, C.C., Kieu, T.D., Chou, Y.C.: A High Payload Steganographic Scheme Based on (7, 4) Hamming Code for Digital Images. In: Electronic Commerce and Security 2008 Symposium, pp. 16–21 (2008)
14. Kim, C.: Data Hiding Based on Compression Dithering Images. Studies in Computational Intelligence, vol. 283, pp. 89–98 (2010)

A Test Framework for Secure Distributed Spectrum Sensing in Cognitive Radio Networks*

Mihui Kim, Hyunseung Choo**, and Min Young Chung

Sungkyunkwan University,
300, Cheoncheon-Dong, Jangan-Gu, Suwon, Gyeonggi-Do, 440-746, Korea
iceblueee@gmail.com, {choo, mychung}@ece.skku.ac.kr

Abstract. To reliably detect primary users (PUs) even on the existence of compromised nodes generating forged sensing reports, secure distributed spectrum sensing (DSS) schemes in cognitive ratio networks (CRNs) have been proposed. However, they have the limitation of sensing accuracy for the existence of PU signals due to the absence of exact signal patterns of PUs. It is caused by FCC restriction (no modification of) on PUs, and thus the CRNs cannot communicate with PUs in order to obtain such patterns. To address this challenge, we propose a test framework utilizing primary user emulation signals that can be applied to existing DSS schemes for reinforcing the robustness against forged sensing values. We evaluate our approach via simulation in comparison with the existing scheme. The results show that our approach improves sensing accuracy and fusion speed in the attack case.

Keywords: Secure distributed spectrum sensing, Forged sensing reports, Test framework, Fusion, Cognitive radio networks.

1 Introduction

Recently, to meet the increasing demand for wireless bandwidth, cognitive radio technologies have been researched. Accurate spectrum sensing in the cognitive radio networks (CRNs) is a critical issue to avoid interfering primary users (PUs) communication and to increase the efficiency of channel utilization through sharing resource between PUs and secondary users (SUs) [1, 2, 3]. Distributed spectrum sensing (DSS) has been recognized as a viable solution to increase the accuracy of sensing even in the dynamic wireless characteristics (i.e., fading or shadowing) [4, 5, 6]. However, the goals of the DSS can be collapsed by only one compromised SU in the worst case [7].

To overcome such vulnerabilities on DSS schemes in CRNs, several secure DSS schemes have been proposed [8, 9, 7] . They exploit the similarity of sensing values in close proximity [9], or the accuracy of past sensing reports [8, 7] in

* This research was supported in part by MKE and MEST, Korean government, under ITRC NIPA-2010-(C1090-1021-0008), FTDP(2010-0020727) and PRCP(2010-0020210) through NRF, respectively.

** Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 382–391, 2011.

order to mitigate the effect of forged sensing reports. However, they do not always provide the high accuracy of spectrum sensing because they are based on estimated PU signal patterns without exact ones. It is induced by the restriction of Federal Communications Commission (FCC); FCC specifically states that "no modification to the incumbent system (i.e., primary user) should be required to accommodate opportunistic use of the spectrum by secondary users" [10].

In this paper, under such a restriction, we propose the second best to test the suspicious SUs generating different sensing results from neighbor SUs or the final decision, with the exact test signal patterns; the test signals are emulated with PU features (i.e., pilot, field sync, segment sync and so on) [11] and are controlled to be received with similar power to one from real PU signals, called as primary user emulation signal (PUES). This PUES has originally received attention as a crucial attack in CRNs. However, we exploit the PUES to test the suspicious SUs on the contrary.

This paper makes the following main contributions:

- Proposal of a general test framework via PUES that can reinforce existing (secure) DSS schemes. It is performed only in the existence of suspicious SUs in order to extract the compromised or malfunctioning SUs.
- Evaluation of our approach through simulation. We instantiate the proposed test framework based on a secure DSS scheme, called as RDSS, and then compare the test scheme with the original RDSS. The simulation results show that our test scheme outperforms the RDSS in terms of both defense performance and fusion speed.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 clarifies network and attack models. Section 4 describes the generic test framework exploiting PUES for the detection of compromised SUs in DSS schemes. Section 5 presents our simulation evaluation of the proposed scheme, and finally Section 6 concludes this paper.

2 Related Work

2.1 Limitation of SDSS

For the success of CRNs, spectrum sensing is the first step but very crucial. However, a single SU as a radio device, may suffer severe shadowing or multi-path fading with respect to PU signal so that it may not detect the existence of PU transmitter even in its vicinities. Thereby, cooperative approaches have been proposed [4, 5, 6]; multiple cognitive radios cooperate to reach an optimal global decision (i.e., fusion) by exchanging and combining individual local sensing results. Allowing multiple CRs to cooperate, distributed cooperative sensing can increase the detection probability, reduce the detection time, and achieve the diversity gain. However, from the view of security, cooperative approaches have a decisive vulnerability, i.e., the threats by compromised nodes [12, 13]. The

threats cannot be solved by basic security mechanisms, i.e., authentication or cryptography. In the worst case, one compromised node may mislead the CRNs [7]; while the node just reports the opposite values to original sensing values.

Therefore, recently several secure DSS schemes has been proposed to enhance the robustness against such vulnerabilities [8, 9, 7]. RDSS [8] mitigates the effect of forged sensing reports through decreasing the weight (called as reputation) while fusing received sensing results to a final decision. However, the mitigation rate is too late, and thus the effect of attack (i.e., causing the interference to PU or decreasing the efficiency of channel utilization) may remain for a quite long time. ADSP [9] explicates the shadowing effect to catch the similarity of sensing results among SUs in close proximity (cluster). In ADSP, the weight as to the similarity degree is imposed in fusion, and moreover if the similarity is too small, the SU is excluded in fusion. However, ADSP has the weakness against major compromised nodes in a cluster (e.g., in the attack case with more than 1/3 compromised nodes). SCSS [7] also uses suspicious level and trust consistency of SUs according to the accuracy of past reports, and thus excludes the SUs with high suspicious and inconsistency levels in fusion. Priori probability for PU signal patterns is assumed in SCSS, but the exact patterns are hard to be obtained from empirical data.

All of these efforts to provide the secure DSS process in CRNs are caused by FCC restriction; FCC prohibits any modification to PUs as mentioned in Section 1. Consequently, any techniques to inform CRNs of PU signal patterns cannot be used directly; none of secure DSS schemes cannot be based on real PU signal patterns, and thus they have the limitation of estimation for such patterns.

2.2 RDSS

In this subsection, we briefly introduce RDSS [8] that we use to concrete the proposed framework. In RDSS, a primary transmitter and an ad-hoc CR network consisting of secondary nodes are assumed. Each secondary acts as a sensing terminal that conducts local spectrum sensing, and then reports their results to a data collector or fusion center. The data collector fuses the received reports to a global decision. In such DSS process, RDSS is proposed to solve the Byzantine failure problem; reporting the falsified local spectrum sensing results.

To address the challenge, RDSS provides a fusion process decreasing the effect of forged sensing reports. Basically it builds the propagation model as log-normal shadowing path loss model, and then it estimates the received power from PU signal for each SU via the model and the location of SU. Finally it performs a weighted sequential probability ratio test (WSPRT) according to received reports and reputation (the accuracy of past sensing reports). Thus, WSPRT for the fusion gradually decreases the influence of compromised SUs through decreasing the reputation of SUs with the sensing value different from the final decision.

3 Network and Attack Models

Our system consists of PUs and CR networks with a base station (BS) and SUs as shown in Figure 1. A PU is assumed to be at a fixed location (e.g., a TV broadcast tower). We assume that SUs are equipped with wireless radio devices and are allowed to transmit signals on the channels allocated to PUs only when the PUs are not transmitting. To protect the primary signal and increase the accuracy of spectrum sensing, each SU conducts spectrum sensing for a periodic quiet period, and gives their sensing results to a center node (referred to as BS hereafter) in a CRN. Then, the BS fuses the received results to a final result, and uses the final result for scheduling the communication channels in the CRN (DSS process). We assume that the BS knows the location of each SU.

The objective of an attacker is to destroy the DSS process, and thus prevent SUs from using the channel of the PUs or cause the interference to the communication of PUs. For that, the attacker may compromise the SUs to forge sensing results, i.e., transmit results that are opposite to its local spectrum sensing results, called as spectrum sensing data falsification (SSDF) attack. Figure 1 depicts an example of gathering the sensing results from SUs including a compromised SU when PU signal exists ('1'). However, a normal SU may report a wrong sensing result ('0') because of the shadowing effect.

Moreover, we assume there exists an underlying basic security framework that enables node authentication and message encryption, and thus eavesdropping, injecting, or altering the sensing results by outside attackers, could be prevented by the security framework. The physical layer is protected using jamming-resistant techniques, such as frequency hopping spread spectrum (FHSS) or direct sequence spread spectrum (DSSS).

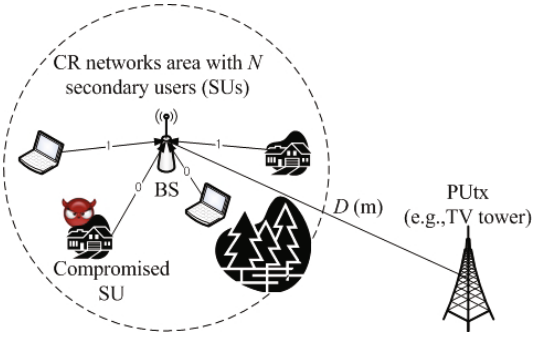


Fig. 1. An example of distributed spectrum sensing with a compromised SU

4 PUES-Based Test Framework

Our goal is to enhance the robustness against SSDF attacks, given the restriction of FCC on (no modification of) PU. As discussed in the Introduction, any DSS

schemes cannot detect the existence of PU signal with high accuracy in all cases due to no help of PU (i.e., giving the signal patterns of PU to CRNs). Thus, we provide a PUES-based test framework that is carried out only if suspicious SUs exist. This framework can be applied to any fusion mechanisms to reinforce the security in DSS.

For the test framework, we define a *test period* that is composed of one or several sensing rounds after recognizing the existence of suspicious SUs, as shown in Figure 2. During the test period, the final decisions follow the sensing results of a BS (u_0) and the BS transmits the PUESs with a test probability P_t to test suspicious SUs during a sensing time when $u_0 = 0$. The test framework with the test period and the general sensing period consists of following steps:

1. Testing nodes in *TestSet* for a test period, if suspicious SUs exist
   (a) Generating PUES for test
   (b) Judging sensing results for test signal
2. Performing distributed spectrum sensing for a general sensing period
   (a) Determining the final result with gathered sensing results
   (b) Composing *TestSet*

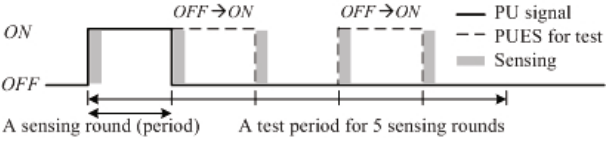


Fig. 2. An example of a test period (e.g.,m=5 and P_t=0.5)

4.1 Testing Nodes in *TestSet* for a Test Period

As shown in Figure 3, a test period is invoked by sensing suspicious SUs, i.e., *TestSet*, a set including such SUs, is not empty. The length of test period consisting of m sensing rounds can be determined by the intended test accuracy; the longer test period is, the more accurate test result outputs but the more overhead of PUES signals is required.

After starting a test period, the BS senses PU signal at each sensing time during the test period, and it then generates PUES with a test probability P_t in order to test the nodes in *TestSet*, only when the BS senses the absence of PU signal (u_0=0). During the test period, only a spectrum sensing result by the BS is used as the final decision for channel scheduling. The BS also records the test information, e.g., sensing times generating PUESs, and its own sensing values for each sensing time.

After gathering sensing results from suspicious SUs in *TestSet* during the test period, the BS evaluates the sensing results comparing with the recorded test information. If the accuracy of the results is below β, the SU is excluded from the next spectrum sensing. The accuracy is calculated as Eq. 1:

$$a_i = \frac{\sum_{j=1}^{m} (-1)^{u_{ij} \wedge (u_{0j} | T_j)}}{m}, \tag{1}$$

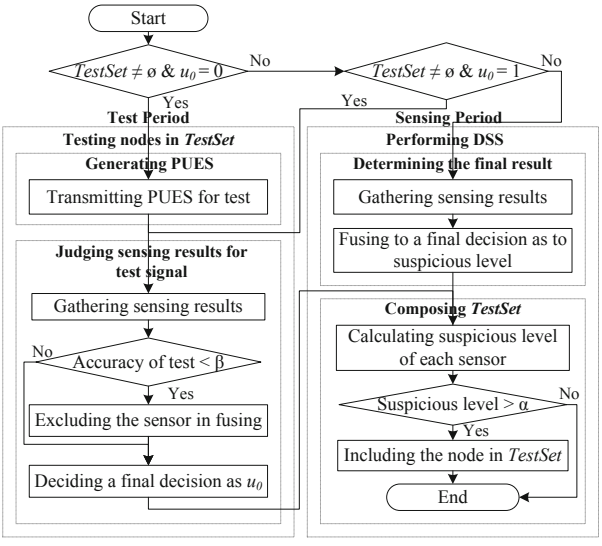


Fig. 3. Flowchart of test framework

where a_i is the test accuracy for node i, u_{ij} and u_{0j} are the sensing result of node i and BS respectively in j^{th} sensing time, T_j is the existence ('1') or absence ('0') of test signal in j^{th} sensing time, and '$\wedge$' and '|' are 'XOR' and 'OR' bit operations.

4.2 Composing *TestSet* for a Sensing Period

If the *TestSet* is empty, the existing DSS schemes [8, 5, 6] are applied. The BS gathers sensing results from SUs after a sensing time and fuses the gathered values to a final decision (the existence or absence of PU signal).

However, after finishing the general distributed spectrum sensing, the second sub-process, composing *TestSet*, should be added in order to judge whether suspicious nodes exist or not. This sub-process is accomplished through examining sensing accuracy of each SU. If suspicious level of any SU is bigger than the predefined parameter α (i.e., the sensing accuracy is low), the SU is included to *TestSet*. The ways to calculate suspicious level can be obtained from existing secure DSS schemes. For example, the suspicious level of SUs can be induced from the similarity among the sensing values of SUs in a cluster [9], or the reputation and suspicious value (related with past sensing accuracy) of each SU [8, 7].

4.3 Instantiation of Test Framework

In this subsection, we briefly present an example to make a concrete test scheme (named as tRDSS) based on an existing secure DSS scheme (RDSS). In RDSS, the reputation value of each SU indicates the accuracy of sensing history. However, if a predefined vale of the reputation is used for composing the *TestSet*

with suspicious nodes, the compromised nodes may give the effect of attack to DSS process for a quite long time until the reputation goes down to the predefined value. Thus, we utilize the variation rate of reputation over a moving window of size w, *trend*. The past w reputation values of node i at time t_n are $r_i^{t_j}(j = n - w + 1, n - w + 2, \ldots, n)$, and thus the trend is calculated as Eq. 2:

$$t_i^{t_n} = \frac{1}{w} \sum_{j=n-w+1}^{n} r_i^{t_j} - r_i^{t_{j-1}}. \tag{2}$$

The calculated trend $t_i^{t_n}$ of node i at time t_n is compared with the predefined threshold α' (≤ 0). If $t_i^{t_n} < \alpha'$ (i.e., it is a decreasing or fluctuated trend), the BS includes the node i to *TestSet* to test the node in the next sensing period. The proper threshold α' can be set up after investigating the trend values in normal case. We will show the distribution of trend values in the next section.

5 Performance Evaluation

In this section, we evaluate the proposed approach through MATLAB-simulation. Our evaluation is focused on the ***defense capability*** (correct sensing ratio (CSR)), miss detection ratio (MDR)), the ***fusion speed*** (number of WSPRT execution), and ***overhead*** introduced by the test framework.

5.1 Simulation Environments

In order to exactly compare tRDSS with RDSS, we configure the simulation network with almost the same parameter values as those of RDSS. In the simulations, 500 secondaries are randomly located in a 2000m×2000m square area, and they form a CR network. The 30% compromised secondaries randomly selected report the opposite values to the original sensing from 10^{th} sensing time. A PU with a duty cycle of 0.2, is located D meters away from the center of the CR network as shown in Figure 1. A secondary acts as a sensing terminal and a BS as a fusion node in DSS performs the test scheme tRDSS if it needs. The interval of sensing period is 30s and each simulation lasts for two hours (refer [8] for more details). However, we set up the transmitted power of PU as a realistic value (100W) according to the distance between BS and PU (3000-6000m) in simulation [14]. In tRDSS, we simply configure both m and β as 1.

5.2 Simulation Results

We carried out the simulations in two attack cases: a normal case and an attack case. In normal case, we varied the distance D. At the far distance as shown in Figure 4, sensing performances (MDR and CSR) decrease a little, as the accuracy of sensing decreases. The number of WSPRT execution required until reaching the final decision increases in two schemes similarly due to the decreasing sensing accuracy as shown in Figure 5.

In the attack case, the results show that tRDSS outperforms RDSS greatly in terms of defense performance and fusion speed. Especially, in RDSS, MDR is high even at the near distance as shown in Figure 6. The reason is analyzed as the effect of 30% compromised nodes is bigger due to high sensing accuracy than the effect at the far distance. Thus, CSR at the near distance is also low due to high MDR. When D is larger than 5000, CSR in tRDSS also decreases due to low sensing accuracy and insufficient sensing reports in the vicinity. However the decrese of CSR in tRDSS can be alleviated through the test framework in comparison with RDSS. Figure 7 shows the fusion speed in tRDSS is considerably fast because most suspicious SUs (attackers) may be fast excluded.

Lastly, Figure 8 shows that the distribution of trends for 500 nodes in normal and attack cases. In normal case (D=3000m), the trend values are mainly distributed from 0.6 to 1 (no values less than 0). In normal case with farther distance (D=6000m), the main distribution moves to between 0.4 and 0.8, and nodes with negative values occur due to decreasing sensing accuracy. However in the attack case, many nodes with negative values present, and the nodes are almost compromised nodes. Thus, we set up -0.2 as the threshold α' to compose the $TestSet$ in this simulation.

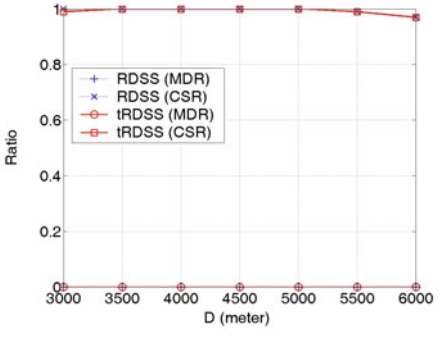


Fig. 4. Ratio vs. distance in normal

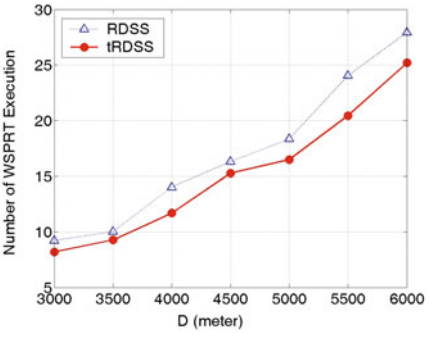


Fig. 5. Fusion speed vs. distance in normal

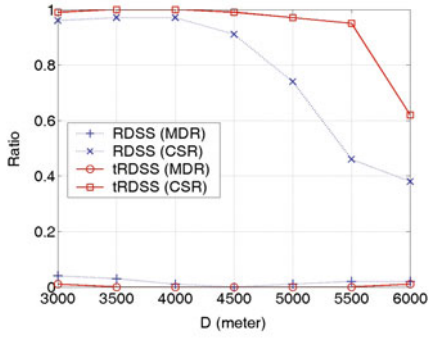


Fig. 6. Ratio vs. distance in attack

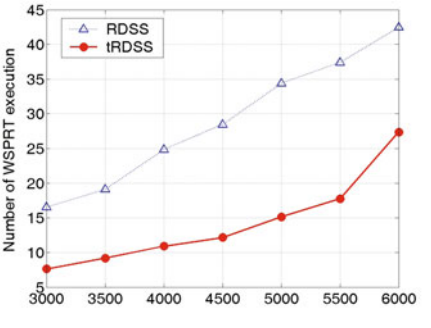


Fig. 7. Fusion speed vs. distance in attack

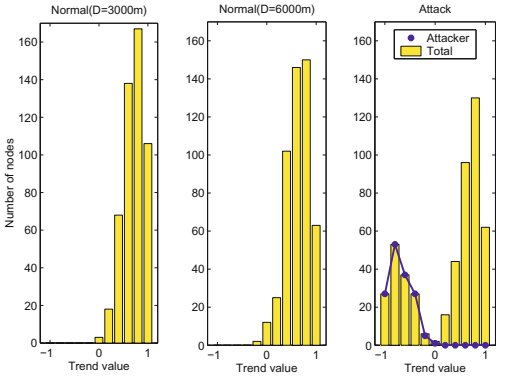


Fig. 8. Trend values in each scenario

5.3 Overhead Analysis

There are two types of overhead introduced by the proposed test framework: computation and communication overheads. However, the computation induced by the trend calculation is performed after finishing each DSS process. Moreover, in the test period, the final decision follows the sensing result of BS. Thus, the extra operations by the test framework do not at all affect the execution time of DSS requiring the fast speed.

In this test framework, PUESs to test suspicious SUs are transmitted, introducing the extra communication overhead $O(m \cdot (1-d) \cdot P_t)$ in a test period, where m is the number of sensing rounds in a test period, d is the duty cycle of PU, and P_t is a test probability (presenting the frequency of PUES). However, the test is performed only when suspicious SUs are sensed.

6 Conclusion

In this paper, we proposed a generic test framework to exclude the compromised nodes on the DSS process in CRNs, in order to overcome limitation of existing DSS schemes under FCC restriction. At first, the test framework composes a *TestSet* with suspicious SUs during the general sensing period. If the *TestSet* is empty, the general DSS scheme is performed continuously. However, if the suspicious SUs exist, the center node (BS) performs PUES test and judges the gathered sensing results of test nodes in *TestSet* in order to distinguish the compromised nodes. Then, we instantiated a concrete scheme called tRDSS based on the proposed test framework and RDSS, an existing secure DSS scheme. To investigate the properties of the proposed scheme, we conducted a set of simulation experiments and compared tRDSS with RDSS. Our simulation results indicate that our approach can successfully enforce a secure DSS scheme to the more secure version. In particular, the instantiated scheme, tRDSS, outperforms previous scheme in terms of both defense and fusion speed performances.

References

[1] IEEE P802.22 Working Group on Wireless RANs, http://www.ieee802.org/22/

[2] Haykin, S.: Cognitive radio: brain-empowered wireless communications. IEEE Journal on Selected Areas in Communications 23(2), 201–220 (2005)

[3] Hossain, E., Niyato, D., Han, Z.: Dynamic Spectrum Accessin Cognitive Radio Networks. Cambridge University Press, Cambridge (2008)

[4] Ganesan, G., Ye, L.: Cooperative spectrum sensing in cognitive radio, part II: multiuser networks. IEEE Trans. on Wireless Comm. 6(6), 2214–2222 (2007)

[5] Ghasemi, A., Sous, E.S.: Opportunistic spectrum access in fading channels through collaborative sensing. Journal of Communications 2(2), 71–82 (2007)

[6] Letaief, K.B., Zhang, W.: Cooperative spectrum sensing. Cognitive Wireless Communication Networks (2007)

[7] Wang, W., Li, H., Sun, Y.L., Han, Z.: Securing collaborative spectrum sensing against untrustworthy secondary users in cognitive radio networks. EURASIP Journal on Advances in Signal Processing 2010 Article ID 695750, 15 (2010)

[8] Chen, R., Park, J.-M., Bian, K.: Robust distributed spectrum sensing in cognitive radio networks. In: INFOCOM, pp. 1876–1884. IEEE, Los Alamitos (2008)

[9] Min, A.W., Shin, K.G., Hu, X.: Attack-tolerant distributed sensing for dynamic spectrum access networks. In: ICNP, pp. 294–303. IEEE, Los Alamitos (2009)

[10] Federal Communications Commission. Facilitating opportunities for flexible, efficient, and reliable spectrum use employing spectrum agile radio technologies. ET Docket (03-108) (December 2003)

[11] Kim, H., Shin, K.G.: In-band spectrum sensing in cognitive radio networks: energy detection or feature detection? In: MobiCom, pp. 14–25. IEEE, Los Alamitos (2008)

[12] Clancy, T., Goergen, N.: Security in cognitive radio networks: threats and mitigation. In: CrownCom (May 2008)

[13] Newman, T.R., Clancy, T.C.: Security threats to cognitive radio signal classifiers. In: Virginia Tech Wireless Personal Communications Symposium (June 2009)

[14] Digital Transmitters Nationwide, http://www.aerialsandtv.com/digitalnationwide.html

The Data Modeling Considered Correlation of Information Leakage Detection and Privacy Violation

Jinhyung Kim and Hyung-jong Kim

Department of Computer Science & Engineering,
Seoul Women's University, Korea
{jinny,hkim}@swu.ac.kr

Abstract**.** Nowadays, the importance of corporations' business information is getting higher and industry people are trying to find software solution preventing the information asset from being disclosed by attackers. There are several representative commercial tools for this purpose and the tools are deployed in many corporations which are handling the critical information such as trade secret, intellectual property and personal information. The tools usually monitor traffic which can contain the important information and also they are watching the e-mail and instant messenger's content. In this work, we are considering the privacy violations in the procedures of data leakage prevention especially the monitoring procedures. In addition, we have tried to make a data model considering the trade-off relation between data leakage prevention and privacy violation. Specifically speaking, we have analyzed the information units of e-mail and instant messenger and assigned a kind of assigned distinct weight values in the privacy and leakage protection viewpoints. In addition, we have shown a case how the weight values are accumulated to represent privacy violation level and data leakage prevention level. Our data model, weight value assignment result and the two kinds of level derivation process are implemented as a database model and user interface.

1 Introduction

Developments in communication technology have made it possible for unauthorized users to access the confidential information of corporations. Such access can create crises that can even threaten the survival of a corporation. The development of communications media such as the Internet has enabled the easy sharing of data; this has led to an increase in the risk of important corporate information being leaked. To mitigate this risk, corporations use systems that are developed to prevent the loss or leakage of information; such systems are referred to as data loss prevention (DLP) systems. However, one consequence of using these systems is that administrators monitoring outgoing data might gain unauthorized access to employees' personal information without their consent. This paper presents a data correlation model for assigning a score to captured packet data and applying the system for DLP. As a main idea, the system collects various network security logs for identification of malicious behavior and correlates the logs. Finally, the system calculates score from result of correlation logs or behaviors.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 392–401, 2011.

Using this model, corporations can protect their data against leakage while at the same time protecting the privacy of their employees.

Section 2 discusses the standards used for the theoretical classification of data according to its degree of importance in terms of DLP Level and according to the degree to which it can compromise employee security privacy level. Further, section 2 presents the proposed theory of data correlation model. In section 3, case studies of private violations caused through the use of DLP systems by corporations are discussed, and possible methods for privacy protection are suggested. Section 3 also presents a data correlation model that assigns a score to out-going data on the basis of its importance in terms of information leakage detection and the degree to which it can compromise employee privacy. Also, various scenarios to which the proposed model can be applied are mentioned and the results of the model's implementation are discussed. Our data model, the two kinds of level derivation process is implemented as a database model and user interface. Section 4 presents some concluding remarks and outline of future research.

2 Trade-Off Model of DLP and Privacy Protection

In this section, the various DLP and privacy levels used to classify information are presented along with their respective definitions. Further, the proposed data correlation model based on the above classification is introduced.

If administrators of a corporation aim for 100% data leakage prevention, they may end up having to tolerate the violation of employee privacy; on the other hand, if they aim for 100% employee privacy protection, they may end up having to tolerate the leakage of confidential data. However, if corporations want to maintain a balance between information leakage and employee privacy protection, the proposed correlation model can be applied.

2.1 Definition of DLP Level

When data packets are filtered through a DLP system, they are classified into four levels on the basis of their degree of importance regarding DLP.

Table 1. Definition of DLP Level

Level	Description
High	Contains sensitive information of a corporation
Medium-High	Data has a direct influence on company's assets
Medium-Low	Data has an indirect influence on company's assets, but (For example, filename or e-mail subject)
Low	Data is not needed for DLP in captured packets

2.2 Definition of Privacy Protection Level

When DLP systems are used, privacy violations can occur while the administrator monitors data flow. To address this concern, the above four levels are applied to privacy protection and are defined in Table 2.

Table 2. Definition of Privacy Protection Level

Level	Description
High	Information distinguishes users directly (e.g., e-mail address) Sensitive personal information
Medium-High	Data can be identified directly who you are or who's behavior by DLP administrator
Medium-Low	Data can be guess who you are or who's behavior by DLP administrator
Low	Information cannot be used to distinguish the sender and/or receiver

2.3 Trade-Off Model

Fig. 1 illustrates the trade-off between the level of importance of data in terms of DLP level and the level of privacy of personal information. Most corporations currently have policies that are configured to maximize only the level of DLP; this inevitably results in privacy violations. However, achieving the optimum balance to address both of these concerns is not as simple as determining the mid-point on a graph, which has been done in Fig. 1. Corporations need to determine the optimum balance by factoring in the nature of the work and the sensitivity of the information that they handle. The correlation model presented in this paper enables corporations to set the level for DLP a required, and to provide detailed data along with a score corresponding to the level of privacy protection.

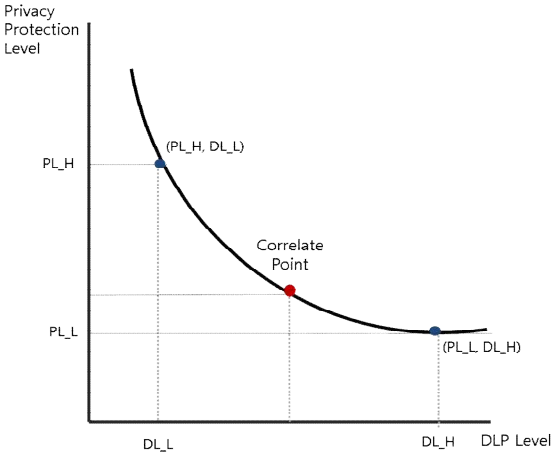


Fig. 1. Correlation of Privacy Protection and DLP Level

3 Data Modeling Considering a Correlation between DLP and Privacy Protection in Email and IM

3.1 Header Analysis of Email and IM (Instant Messenger)

The system in this work involves the construction of a mathematical model that is based on the results of outgoing data classification according to the levels listed in Table 1 and Table 2. The message of email and IM can be represented as the header, body and attached files as shown in Table 3. Table 3 shows composition of each message component.

Table 3. Elements of E-mails and IM Messages [2]

		Element	Type
E-mail	Header	From, To, CC, BCC, Subject, Reply-to, Date, Timestamp	txt
	Body	Content	Txt
	Files	filename	txt, bin
IM	Header	Port Number, Sender Email ID, Receiver Email ID, IP	Txt
	Body	Content	Txt
	Files	filename	txt, bin

Especially, in the header information case, each component can have its own sensitivity in terms of DLP level and Privacy level. The Table 4 shows their DLP level and Privacy level considering the meaning of each element.

Table 4. Levels of Data of Email and IM System Header

DLP Level	E-mail_Data	Privacy Level	DLP Level	IM_Data	Privacy Level
3	To	2	3	SBserver IP	2
3	Cc	2	3	Sender E-mail(ID)	2
3	Bcc	2	3	Port Number	2
4	Subject	1	1	Receiver E-mail(ID)	4
2	From	3			
2	Reply-To	3			
3	Received	2			
1	Message ID	4			
3	Date	2			
1	MIME=Version	4			
1	Content-Type	4			
1	Charset=”euc-kr”	4			

In addition, we have derived a value which represents the header’s DLP level and Privacy level using the expression (1). The aim of the expression (1) is normalization of the header’s DLP level.

$$Header_{dlp} = \frac{\text{Sum of DLP levels of Selected Components}}{\text{Sum of DLP levels of Header's all Component}}$$

$$Header_{pri} = \frac{\text{Sum of Privacy levels of Selected Components}}{\text{Sum of Privacy levels of Header's all Component}} \quad (1)$$

Where, $Header_{dlp}$: DLP level of a Header
$Header_{pri}$: Privacy level of a Header

3.2 Calculation of DLP and Privacy Level

3.2.1 Basic Expression

$$\text{Level}_{dlp} = \alpha \times (\text{Header}_{dlp}) + \beta \times (\text{Body}_{dlp}) + \gamma \times (\text{AttachedFile}_{dlp}) \quad (2)$$

where, $\alpha+\beta+\gamma=1$
$0 \leq \alpha, \beta, \gamma \leq 1$

Equation (2) is a formula used to assign a score of DLP to e-mails and IM's messages. As the expression shows, there are three coefficients which represent the weight of components of the messages. At this moment, we use the fixed three values α, β, and γ for representation of components' contribution to the score, as shown in Table 5. The Level_{pri} is also calculated in the same manner of (2) using same coefficients.

Table 5. Weight of the each element in Email and IM

Classification	Message's Composition	Weight
Header	Sender, receiver, CC(Bcc), Subject, Timestamp, etc.	0.2(α)
Body	Message, Content	0.4(β)
Attached File	OLE, File (PPT, Doc, Txt, Etc.)	0.4(γ)

3.2.2 Case Study

In this section, a case is introduced to present how the suggested expressions are applied in real scenario.

<Scenario>

Corporation A starts to monitor the e-mails sent by one of their employees, Bob, using a DLP system.

1. Corporation A wants to intercept an e-mail sent by Bob (bob@mail.com) to Alice (alice@host.com)
2. When the monitoring is started, the DLP system starts detecting any data being sent from Bob's e-mail address
3. By default, the header information should be confirmed without checking the contents of the text or opening any attachments. Typically, the internal information flow to the body, rather than information listed in the attached file confidential documents attached to the inside if you try to send a frequent.
4. When performing a data check through a monitoring process, the proposed model calculates using the expression (1) and (2).

Table 6. Example of Sending an E-mail

Element	Content	Check
To:	alice@host.com	O
Cc:	Empty	O
Bcc:	Empty	O
Subject	Hello friend!	O
From	bob@mail.com	O
From:""	Empty	O
Reply-To:	bob@mail.com	O
Received:		X
Message_ID		X
Date:	2010-02-17 am 10:16	O
MIME-Version:		X
Content-Type:	Text/Plain, Multipart	O
charset="euc-kr"		X
Body_Check	Hello, how are you doing? Read enclosed document. It's between ourselves	X
AttachedFile_Check	Title: The employee's salaries: Will Smith – 500$ Kate Brown – 600$ Andy Hue – 360$	O

Table 6 is an example of captured message contents of e-mail from Bob to Alice. In this case, it checks header information of email like To, Cc, Bcc, Subject, From, Reply-to, Date, Content-Type and AttachedFile's name of captured packets. The calculation results of DLP level and Privacy level are as follows.

[DLP Level]

$$Header_{dlp} = DL_{To}(3) + DL_{Cc}(3) + DL_{Bcc}(3) + DL_{Subject}(4) + DL_{From}(2) + DL_{Reply-To}(2) + DL_{Date}(3) + DL_{Content-type}(1) \quad /27$$

$$= 21/27$$

$$= 0.78$$

$$Body_{dlp}=0$$

$$AttachedFile_{dlp}=1 \qquad \ldots ①$$

According to result of ①, the DLP is calculated as shown in below;

$$DLPLevel = 0.2 \times (Header_{dlp}) + 0.4(Body_{dlp}) + 0.4(AttachedFile_{dlp})$$

$$=0.2\times 0.78 + 0.4 \times 0 + 0.4 \times 1$$

$$\mathbf{=0.56} \qquad \ldots ②$$

[Privacy Level]

$$Header_{pri} = PL_{To}(2) + PL_{Cc}(2) + PL_{Bcc}(2) + PL_{Subject}(1) + PL_{From}(3) + PL_{Reply-To}(3) + PL_{Date}(2) + PL_{Content-type}(4) \quad /33$$

$$=19/33$$

$$= 0.58$$

$$Body_{pri}=1$$

$$AttachedFile_{pri}=0 \qquad \ldots ③$$

According to result of ③, Privacy Level is calculated as shown in below;

$$PrivacyLevel = 0.2 \times (Header_{pri}) + 0.4(Body_{pri}) + 0.4(AttachedFile_{pri})$$
$$= 0.2 \times 0.58 + 0.4 \times 1 + 0.4 \times 0$$
$$= \mathbf{0.516} \quad \ldots ④$$

Based on the above results like ①, ②, ③, ④, the DLP Level is slightly higher than the Privacy Level.

The proposed formula can be applied to eight possible scenarios as shown in Tables 7 and 8. We hypothesized an inverse relation between the DLP level and privacy protection level, and this hypothesis is proved Fig. 2 by plotting the graph of the data in Tables 7 and 8.

Table 7. Using the three elements to create a case

	1	2	3	4	5	6	7	8
Header	O	O	O	O	X	X	X	X
Body	O	O	X	X	O	O	X	X
Attached File	O	X	O	X	O	X	O	X

Table 8. Result of each calculated case

	H(0.2)_D	B(0.2)_D	A(0.4)_D	**D_Sum**	H(0.2)_P	B(0.4)_P	A(0.4)_P	**P_Sum**
1	0.71	1	1	**0.942**	0.12	0	0	**0.024**
2	0.71	1	0	**0.542**	0.12	0	1	**0.424**
3	0.71	0	1	**0.542**	0.12	1	0	**0.424**
4	0.71	0	0	**0.142**	0.12	1	1	**0.824**
5	0	1	1	**0.8**	1	0	0	**0.2**
6	0	1	0	**0.4**	1	0	1	**0.6**
7	0	0	1	**0.4**	1	1	0	**0.6**
8	0	0	0	**0**	1	1	1	**1**

(where, H_D = $Header_{dlp}$ level, H_P = $Header_{pri}$ level, B = Body level, A = Attachedfile level, D_Sum =sum of H_D, B, A, P_Sum = sum of H_P, B,A)

3.3 Implementation Results

As a concept proving implementation of our suggesting trade-off model and calculation method, we have implemented a database model and web based user interface. Fig. 3 shows the policy editing view which enables the administrator to choose the monitoring target for DLP and Privacy Protection. Based on the selected items, the DLP level and Privacy Protection level can be calculated and saved into the database. The Fig. 4 shows an web-based user interface which will be shown to administrator.

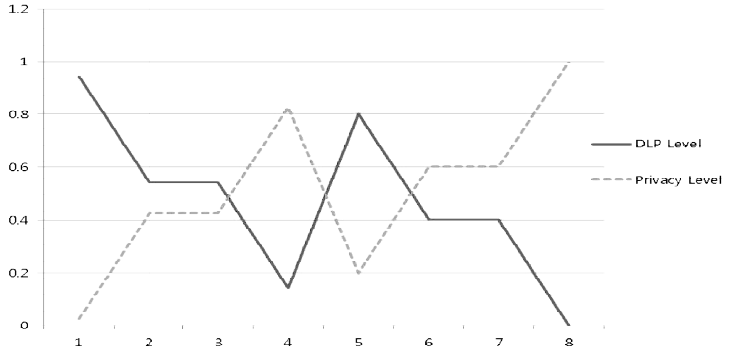


Fig. 2. Correlation of DLPLevel and Privacy Level

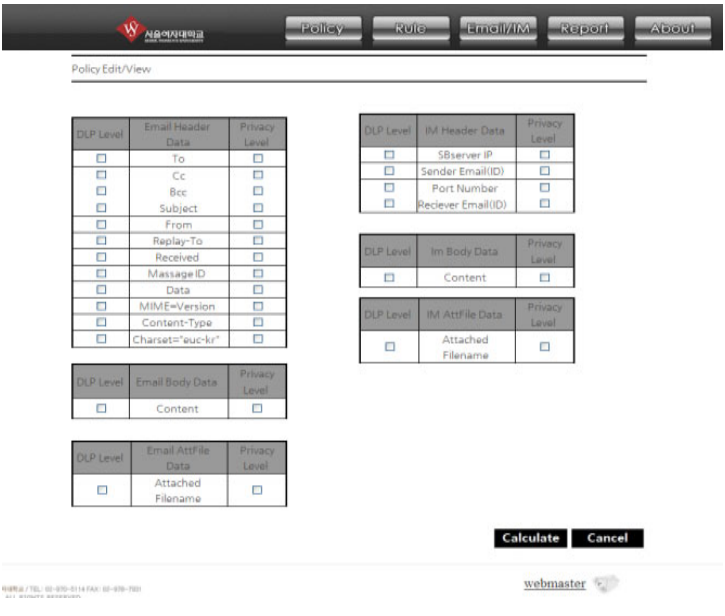


Fig. 3. Policy Edit/View

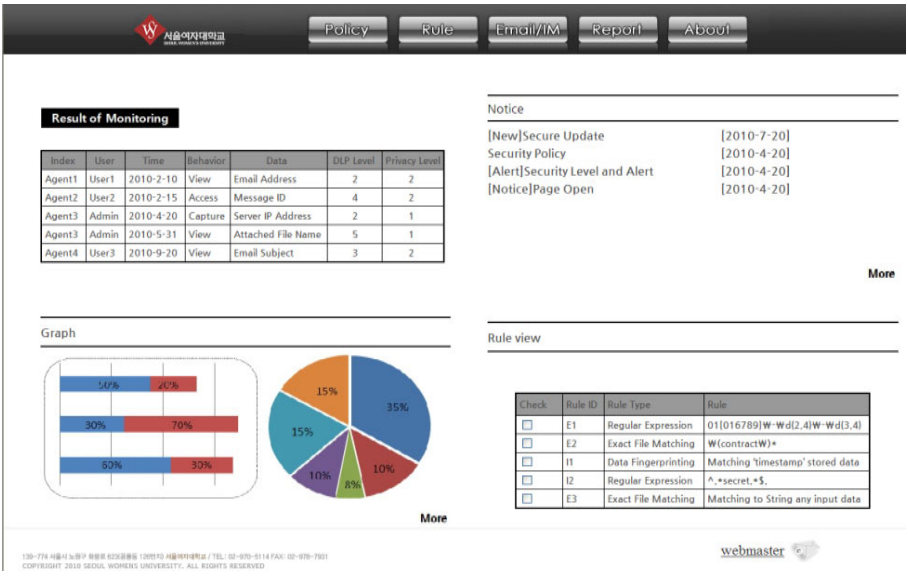


Fig. 4. Results of Implementation

4 Conclusion

In this paper, information regarding the sensibility of both personal and internal-corporation information is analyzed, along with a model that offers a data score that detects a trade-off model between the two sensibilities. Further, a concept proving implementation is presented to rank the results of collected messages of e-mail and IM obtained using the proposed model. This model can also be used to define privacy protection policies within corporations.

The fact that the developed model discussed in this paper can be used to detect information leakage and to protect privacy according to effective inner-corporation policies and at different levels is of great significance.

Acknowledgement

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (2009-0068361).

References

1. Jia, X.-P., Peng, H., Zheng, Q.-L., Jiang, Z.-L., Li, Z.: A Topic-Based Document Correlation Model. In: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, July12-15 (2008)

2. Kim, J.H., Kim, H.J.: Design of Internal Information Leakage Detection System Considering the Privacy Violation. In: ICTC2010 (International Conference on ICT Convergence 2010) (2010)
3. Stolfo, S.J., Hershkop, S., Hu, C.-W., Li, W.-J., Nimeskern, O., Wang, K.: Behavior-Based Modeling and Its Application to Email Analysis. ACM Transactions on Internet Technology 6(2), 187–221 (2006)
4. Choi, D., Jin, S., Yoon, H.: A Personal Information Leakage Prevention Method on the Internet, 3rd edn. Springer, Heidelberg (1996)
5. Liu, S., Kuhn, R.: Data Loss Prevention, vol. 12(2), pp. 10–13. IEEE Computer Society, Los Alamitos
6. Gómez-Hidalgo, J.M., Martín-Abreu, J.M., Nieves, J., Santos, I., Brezo, F., Bringas, P.G.: Data Leak Prevention through Named Entity Recognition. In: 2010 IEEE Second International Conference on Social Computing, SocialCom, August 20-22, pp. 1129–1134 (2010)
7. Hooper, E.: Intelligent strategies for secure complex systems integration and design, effective risk management and privacy. In: 3rd Annual IEEE Systems Conference, March 23-26, pp. 257–261 (2009)

A* Based Cutting Plan Generation for Metal Grating Production

Jin Myoung Kim and Tae Ho Cho

School of Information and Communication Engineering,
Sungkyunkwan University,
Suwon, Kyungkido, South Korea
{kjm77,taecho}@ece.skku.ac.kr

Abstract. In the metal grating production process, a cutting plan should decide how pieces of metal rectangles are allocated and cutout from plate sheets called panels. The cutting plan can generate various possible combinations of rectangle allocations within the panels in order to select the best plan that minimizes material waste. To achieve the best plan, A* algorithm of artificial Intelligence is exploited. The plan is evaluated to show how effective it is in terms of material utilization.

Keywords: A* algorithm, cutting plan, grating production, combinatorial optimization.

1 Introduction

In modern manufacturing industries, industries, automatic systems composed of computers are common. The CAM (Computer aided manufacturing) is defined as the effective use of computer technology in the planning, management, and control of the manufacturing function [1]. The cutting problem is NP-complete with a number of industrial and commercial applications [2]. This problem appears in the cutting of steel plates into required sizes, in the cutting of wood sheets to make furniture and in the cutting of cardboard into boxes. The related problem of minimizing the amount of waste, or a loss produced by the cutting can be converted into this problem by making the value of all pieces equal to their areas [3].

The cutting stock problem has been treated first by Kantorovich [4] and later by Gilmore and Gomory [5,6]. Christofides and Whitlock [7] proposed a depth-first branch-and-bound algorithm in which all possible guillotine cutting patterns are enumerated by constructing a tree[2], as does Beasley with the nonguillotine variant [8]. However, once again, with larger instances the method becomes time infeasible. Beasley compares both optimal and heuristic algorithms using dynamic programming [3]. Hifi and Zissimopolous [9] presented an exact algorithm that improves on the approach used by Christofides and Whitlock [10].

Recently, Van Dat Cung et al [11] developed a new version of the algorithm proposed in Hifi and Zissimopolous [9] that used a best-first branch-and-bound approach to solve exactly some variants of cutting problem [10]. Viswanathan and Bagchi [2]

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 402–410, 2011.

applied the best-first search method to solve two-dimension cutting stock problem, and Victor Parada Daza *et al*. [12] improved Wang's algorithm [13] by heuristic method.

Another method for placing rectangles on the plates is BL (Bottom Left) method in which rectangles are allocated based on the first come first serve principle [13,14]. Chazelle [15] applied the BL method in finding the lowest possible BL stable location and broke ties by taking the leftmost. Jakobs [16] used a BL method that takes as input a list of rectangles and places each one in turn onto the stock sheet. Liu and Teng [17] offer a new method that allows for representation of any BL stable solution.

The products that are manufactured by the S company near Seoul, South Korea are the metal grating of various rectangle shapes. The company wants to generate an efficient cutting plan in order to reduce loss of material before the actual production process. The cutting plan should decide how pieces of metal gratings are allocated and cutout from the panels. The problem of cutting plan generation in metal manufacturing processes is similar to that of two-dimensional cutting stock problem. This paper deals with the problem of rectangle shaped grating allocations in order to select the best cutting plan that minimizes material loss. To solve this problem, the S company's manufacturing process is identified first and then A* algorithm is applied.

2 Related Work

Some of the research lists in the previous section are further in the following subsections.

2.1 Bottom Left

The size of the search space of the orthogonal packing problem is infinite since every movement of a rectangle into a packing pattern is a feasible direction creates a new packing pattern [16]. The BL heuristic, introduce, in [14], is perhaps the most widely used heuristic for placing rectangles [18].

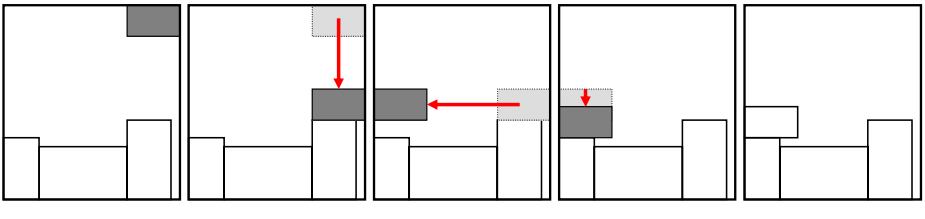


Fig. 1. A bottom-left strategy

This allocation strategy first places the rectangles in the top-right location and makes successive moves of sliding it as far down and left as possible [10].

2.2 Depth-First Breach-and-Bound

Christofides and Whitlock [7] used a depth-first branch-and-bound tree search method to solve a two-dimensional guillotine problem [10]. The algorithm starts with positioning the stock rectangle at the root of the tree. The collections of rectangles that are achieved by a guillotine cutting process on the stock rectangle is represented by the rest of the nodes in the tree. Branching activity of the edges on the tree corresponds to the guillotine cutting process [2].

3 Manufacturing Metal Gratings

Metal gratings, simply defined, are open grid assemblies of metal bars in which the principal load bearing bars run parallel in one direction and are spaced equidistant from each other, either by attachment to cross bars running in a perpendicular direction.

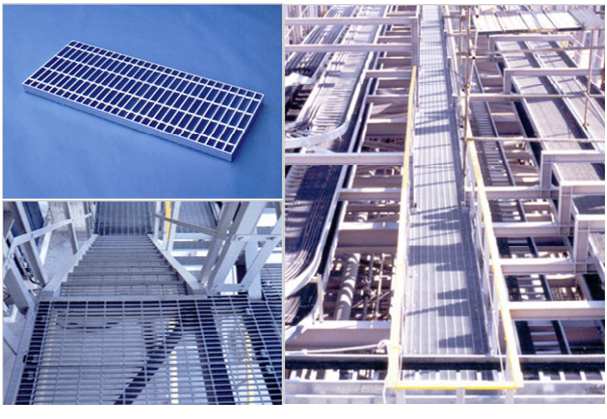


Fig. 2. Grating shape and usage

These gratings are used to cover water drains along the sides of streets. I is also used as a flooring material for various types of production plants and ships. Figure 2 shows a shape and usage of grating.

The manufacturing process of the metal grating is proposed of a requirement analysis, layout design, project welding, cutting and testing. Figure 3 shows a step of a cutting process. Before cutting gratings, manufacturers consider a way how to cut a panel to reduce a material loss. The layout design of gratings on the panel should be considered for reducing manufacturing cost since the material loss is decided in this step. In figure 3, thick lines imply a machine cutting line in the layout sheet. The machine cutting is applied for complete vertical cuttings of a panel. In the slot, dash lines imply a manual cutting line by human. Shade portion implies material loss.

Thus, the cutting plan should be generated in such a way that it should maximize the panel utilization. The cutting plan denotes the layout design.

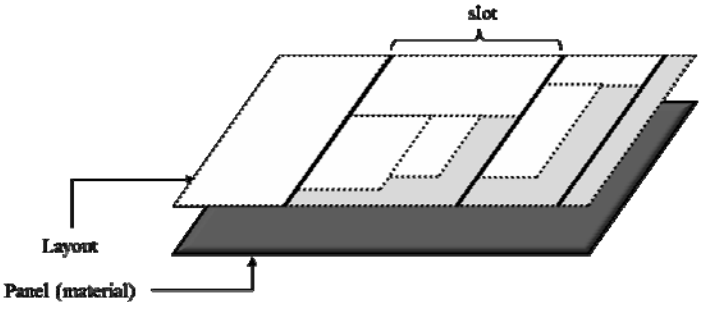


Fig. 3. Cutting type and cutting plan

4 Application of A* Algorithm

4.1 A* Algorithm

A* algorithm, which is a kind of tree search method, uses a heuristic evaluation function [20]. The heuristic evaluation function helps decide which node in the best one to expand next.

Let h(n) be the actual cost of the minimal cost path between node n and a goal node and let g(n) be the cost of a minimal cost path from the start node, n_0 to node n. Then equation (1) is the cost of a minimal cost path from n_0 to a goal node over all paths that are constrained to go through node n.

$$f(n) = g(n) + h(n) \tag{1}$$

For each node n, let heuristic factor $h^*(n)$ be some estimate of h(n), and let depth factor $g^*(n)$ be the cost of the lowest-cost path found by A* so far to node n. The estimate of heuristic evaluation if defined:

$$f^*(n) = g(n)^* + h(n)^* \tag{2}$$

In algorithm A* we use equation (2).

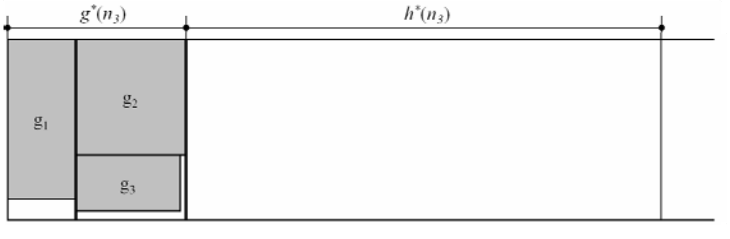


Fig. 4. Hypothetical layouts in a panel

Fig. 4 shows a hypothetical layout of gratings a panel. Grating g_1, g_2 and g_3 are elected by A* algorithm. In tree search space, a root node selects next node from candidate nodes using an evaluation function. In cutting plan, the root node and candidate nodes correspond to a panel and gratings. Also an edge between nodes corresponds to material loss and it is estimated by the evaluation function. Thus, the shortest path found in the tree becomes the optimal plan in the cutting plan in the cutting process.

4.3 Grating Layout

As shown in fig. 3, a machine cutting is applied for the vertical cut which results in a slot. The slot is decided by the grating allocated at the top position of a panel. So, the slot is created by machine cutting only. Manual cutting consists of vertical and horizontal cuttings. Fig. 5 shows the layout of gratings within a slot.

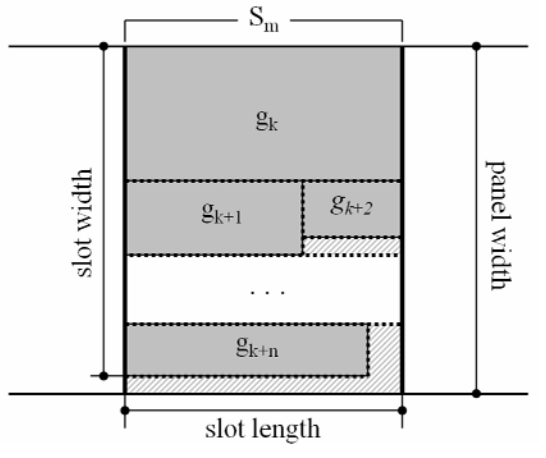


Fig. 5. Layout of gratings within a slot

The sub-gratings are collection of more than on grating that are allocated horizontally within a slot. In fig. 5, g_{k+2} is sub-grating of g_{k+1}.

5 Heuristic Evaluation Functions

5.1 $g^*(n)$: Estimate of Depth Factor

The heuristic evaluation function is shown in equation (2). For an arbitrary node n_c, let S_c be a set of allocated nodes within the cth slot and an element of S_c is s, then the depth factor g^* is defined as follows:

$$g^*(n_c) = g^*(n_{ancestor}) + \text{cost}(n_c)$$

Where

(3)

$$\text{cost}(n_c) = \left[\{PW \times MaxLen(S_c)\} - \sum_{k=1}^{s} (W_k \times L_k) \right] + |n_c|_{cutting}$$

In equation (3), the cost consists of the material loss and the number of cuttings need for n_c. PW is panel width and MaxLen(S_c) is the maximum length of the gratings within slot S_c. W and L is a width and a length of a grating, respectively. So, the left side of the cost function is a material loss in the slot S_c. $|n_c|_{cutting}$ is the number of the cuttings needed for producing the grating n_c.

5.2 $h^*(n)$: Estimate of Heuristic Factor

The $h^*(n)$ represents the future cost regarding the node not yet expanded in a panel. So, $h^*(n)$ is applied in evaluating the result of the hyper-allocation of the gratings to the unallocated portion of the current panel being planned.

Let M' be a set of nodes that exclude node n_c and ancestors of node n_c, i.e., a set of nodes not yet expanded. The hyper-allocation algorithm is shown below:

Hyper-Allocation Algorithm

1. The elements of M' is sorted by the width of elements in descending order.
2. The first element in the sorted list of M' is allocated and is removed from M'. If there is no element in the M', exit algorithms.
3. Loop:
   3.1 In M', if there is an element which can be allocated in the slot formed by 2 then the element is allocated and is removed from M'.
   3.2. If the number of elements of M' is zero, then exit the loop.
4. In M', if there is no element which can be allocated in the slot, then go to 2.

Grating allocation with the hyper-allocation algorithm does not ensure optimal layout of gratings. So, the material loss for the algorithm should be defined. Let S' be a set of slots generated by the algorithm and the number of elements in S' is s, then $h^*(n_c)$ is defined as:

$$h^*(n_c) = \sum_{k=1}^{s} (PW - SW_k) \times SL_k + \sum_{k=1}^{s} |n_k|_{cutting} \quad (4)$$

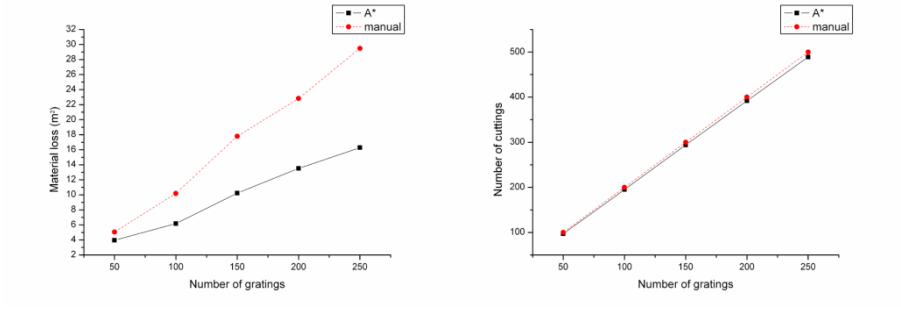
In equation (4), SW and SL are a width and a length of a slot, respectively. The first term of the equation is the sum of the total expected loss for the remaining allocation

in a panel. The equation only designates a material loss for SW is less or equal to 80% of PW. This is a reasonable criterion by export cutting planner in the S company.

6 Simulation Results

Simulation is performed to show the effectiveness of the proposed cutting algorithm against the current manual plan generation method adopted by expert cutting planners in S company. The expert cutting planners seldom look for the adaptively grating allocation due to the complexity in the cutting plan. Their general approach in the cutting plan is based on a simple greedy algorithm. A wider grating has higher priority than narrow one in the allocation. That is, the expert cutting planner firstly allocates the widest grating on the panel.

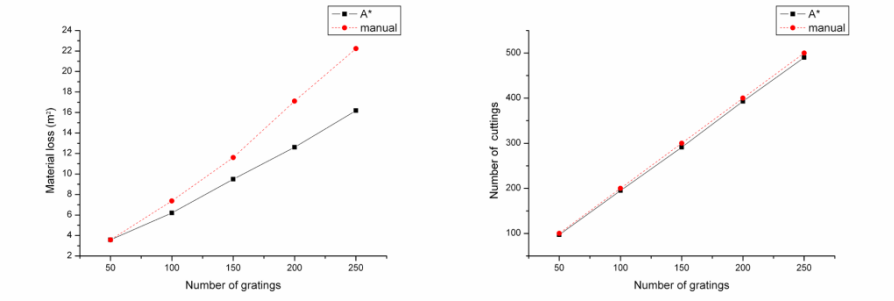
The simulation is performed on two different cases. For both cases we observed the amount of loss of raw material and number of cuttings needed in production. In the first case, the gratings to be produced are randomly decided so that we can have as many different types of gratings as possible. The gratings to be produced in the second case are decided based on the last one year of real production data of S company. There are relatively more identical types of gratings compared to the first case. The dimensions of the gratings in both cases range from 300mm to 995mm for both sides of rectangles. About 30% of the gratings are of identical size in the second case.



(a) (b)

Fig. 6. Evaluation of material loss (a) and number of cuttings (b) for randomly generated gratings

Figure 7 (a) shows that the material loss in the grating production with the A* algorithm is much less than the manual plan generation method. As the number gratings to be produced are increased, the difference in material loss becomes larger due to the increased complexity in the plan generation. The number of cuttings needed in grating production is shown in figure 7 (b). It shows that efficient use of the material in the A* algorithm is achieved without the added cost of cutting operation.



(a) (b)

Fig. 7. Evaluation of material loss (a) and number of cuttings (b) for gratings produced by S company

Figure 8 (a) and (b) show the simulation results of the material loss and count of the number of cuttings in grating production, respectively. The difference in material loss is slightly less in this case and the rest of the simulation results are similar to that of the first case. As shown in figure 8 (a) the difference between the A* algorithm and the manual method is about 6 m2 in producing 250 gratings. If a company produces 10,000 gratings per year, the difference will be about 240 m2. For production of 20,000 gratings per year, it will be 480 m2. This represents a significant reduction in material waste. The difference tends to be larger in the current actual grating production environment since the frequent plan modifications are requested in the middle of the grating production and not all these requests can be accepted due to the slow manual plan generation.

7 Conclusion

Cutting plan generation is one of the most important parts of grating production since the plan determines the amount of material loss. Generally the material loss affects mostly the grating production cost. The existing manual plan generation method is slow and inefficient and the number of employees that participate in the plan generation is larger than in a computerized system.

We proposed a computerized plan generation system implemented based on the A* algorithm, in which each grating is represented as a node to be expanded in the search tree. As described in this paper, the A* algorithm applies well in solving the problem of grating plan generation. The simulation results demonstrate that the proposed A* algorithm based cutting plan is much more efficient, in terms of material usage, without the added cost of cutting operations, over the manual cutting plan generation method. Further, the computerized system enables fast plan generation when plan modifications are requested. As a result, the overall production time can also be reduced. The future research includes refinement of the algorithm by considering additional factors such as the gap between gratings, blade type of machine saw, and so on.

References

1. Groover, M.P.: Automation, Production system, and Computer Integrated Manufacturing. Prentice-Hall, Englewood Cliffs (1987)
2. Viswanathan, K.V., Bagchi, A.: Best-first search methods for constrained two-dimensional cutting stock problems. Operations Research 41, 768–776 (1993)
3. Beasley, J.E.: Algorithms for unconstrained two-dimensional guillotine cutting. Journal of the Operational Research Society 36, 297–306 (1985)
4. Kantorovich, L.V.: Mathematical method of organizing and planning production. Management Science 6, 363–442 (1960)
5. Gilmore, P.C., Gomory, R.E.: Multistage cutting stock problems of two and more dimensions. Operations Research 13, 94–120 (1965)
6. Gilmore, P.C.: Cutting stock, linear programming, knapsacking, dynamic programming and integer programming, some interconnections. Annals of Discrete Mathematics 4, 217–235 (1979)
7. Christofieds, N., Whitlock, C.: An algorithm for two dimensional cutting problems. Operations Research 25, 30–44 (1977)
8. Beasley, J.E.: An exact two-dimensional non-guillotine cutting tree search procedure. Operations Research 33, 49–64 (1985)
9. Hifi, M., Zissimopolous, V.: Constrained two-dimensional cutting: An improvement of Christofides and Whitlock's exact algorithm. Journal of Operation Research Society 5, 8–18 (1997)
10. Burke, E.K., Kendall, G., Whitwell, G.: A New Placement Heuristic for the Orthogonal Stock-Cutting Problem. Operation Research 52, 655–671 (2004)
11. Cung, V.-D., Hifi, M., Le Cun, B.: Constrained two-dimensional cutting stock problems best-first branch-and-bound algorithm. Operations Research 7, 185–210 (2000)
12. Daza, V.P., et al.: Exact solutions for constrained two-dimensional cutting problems. European Journal of Operations Research 84, 633–644 (1995)
13. Wang, P.Y.: Two algorithms for constrained two-dimensional cutting stock algorithms. Operations Research 31, 573–586 (1983)
14. Baker, B.S., Coffman Jr., E.G., Rivest, R.L.: Orthogonal packings in two dimensions. Society for Industrial and Applied Mathematics Journal of Computing 9, 846–855 (1980)
15. Chazelle, B.: The bottom-left bin packing heuristic: An efficient implementation. IEEE Transaction on Computers 32, 697–707 (1983)
16. Jakobs, S.: On genetic algorithms for the packing of polygons. European Journal of Operations Research 88, 165–181 (1996)
17. Liu, D., Teng, H.: An improved BL-algorithm for genetic algorithms of the orthogonal packing of rectangles. European Journal of Operations Research 84, 539–561 (1999)
18. Lesh, N., Marks, et al.: New Heuristic and Interactive Approaches to 2D Rectanglar Strip Packing. Journal of Experimental Algorithmics 10, 1–18 (2005)
19. Nilsson, N.J.: Artificial Intelligence: A new synthesis. Morgan Kaufmann Publishers, San Francisco (1998)

Intelligent Forecasting of S&P 500 Time Series — A Self-organizing Fuzzy Approach

Chunshien Li and Hsin Hui Cheng

Laboratory of Intelligent Systems and Applications
Department of Information Management
National Central University, Taiwan, R.O.C.
jamesli@mgt.ncu.edu.tw

Abstract. Stock index time series may allow investors to become aware of the change of stock market. In the paper, we aim at forecasting S&P 500 Index, one of the most representative stock indices in United States. A self-organizing fuzzy-based approach for intelligent predictor is used. The design for the predictor is divided into the structure and parameter learning stages. The FCM-Based Splitting Algorithm is used to determine the optimal number of fuzzy rules for the predictor. Two hybrid learning algorithms, the PSO-RLSE and PSO-RLSE-PSO methods, are used for the parameter learning of the predictor, respectively. To test the proposed approach, we devise experiments to compare the performances by the intelligent predictor trained with the two learning algorithms, respectively. Moreover, an additional experiment for different input orders is conducted to see the influence on the performance. The excellent performances in accuracy by the proposed intelligent approach are exposed.

Keywords: Forecasting, Hybrid Learning, Clustering, Fuzzy System, Particle Swarm Optimization (PSO), Recursive Least Squares Estimator (RLSE).

1 Introduction

System modeling is usually based on systemic thinking and is important to various research areas, such as forecasting [1], control, signal processing, and many others. System modeling can provide an interactive environment for designers to test and confirm a specific part of an overall problem. Time series is a sequence of data points which are measured typically at successive time intervals, for instance, the stock index. Stock indices can make investors to understand the changes of stock market. The change of stock market means opportunity and risk of economic arrangement. We choose the S&P 500 index [2] as a research subject. For many large-cap American stocks, the S&P 500 is one of the most widely influential indices. Published since 1957, the S&P 500 is an index of the prices of the most representative five hundred companies in NYSE Euronext and Nasdaq OMX group. And, it is calculated by a free-float capitalization-weighted method [3]. Many kinds of funds, such as pension funds and mutual funds, are planned to track the performance of the S&P 500 index.

For time series prediction, there exist a considerable number of studies in literature. Artificial neural network (ANN) is one of commonly used approaches, for which several optimization algorithms have been used [4]-[8]. For example, Luna et al.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 411–420, 2011.

suggested a constructive fuzzy system modeling based on Takagi-Sugeno system and the expectation maximization technique for time series prediction [4]. Chen et al. introduced a new time series forecasting model based on the flexible neural tree (FNT) [5]. Rojas et al. designed a hybrid ANN and ARMA models to solve time series problems [6]. Wang et al. proposed a novel adaptive neural network (ADNN) with the adaptive metrics of inputs and a new mechanism for admixture of outputs for time-series prediction [7].

In this study, we use neural fuzzy theory and clustering method to design the intelligent predictor, and then propose a PSO-RLSE-PSO hybrid learning algorithm to train the predictor for the prediction of the S&P 500. A neural fuzzy system (NFS) [9] [10] is not only a useful model to process linguistic information with uncertainty [11] but also can be greatly practical in function approximation. In this study, the type of Takagi-Sugeno (T-S) fuzzy system [12] is designed and utilized in the aspect of time series forecasting [13] [14]. The design of the intelligent predictor includes two stages, which are the structure learning stage for the structure arrangement of the predictor and the parameter learning stage for the optimal performance by the predictor. The size of the rule base for the neural fuzzy system is automatically determined by a clustering algorithm, called FCM-Based Splitting Algorithm (FBSA) [15]. To adapt the free parameters in the NFS-based predictor, a PSO-RLSE-PSO hybrid learning algorithm with the particle swarm optimization (PSO) [16] and the recursive least squares estimator (RLSE) [12] is used to make the prediction as accurate as possible. The PSO-RLSE-PSO is an improved method from the hybrid PSO-RLSE method. And it can upgrade the prediction performance. In addition, this study also investigates the arrangement of the predictor inputs for its possible impact on the forecasting performance.

In Section 2, the rationale of the study is given, including the theory of neural fuzzy system (NFS), the FBSA clustering method for structure learning, and the PSO-RLSE-PSO hybrid learning method for parameter learning. In Section 3, experiments are devised to test the proposed method for the prediction of the S&P 500. The results by the proposed PSO-RLSE-PSO learning method and by the PSO-RSLE learning method are compared. The experiment for different arrangement of the predictor inputs is conducted in Section 3 as well. A good discussion is given in Section 4. Finally the paper is concluded.

2 Methodology

In this section, we specify the rationale of the proposed approach for the S&P 500 time series forecasting. Firstly, the theory of neural fuzzy system (NFS) is given. The operation of the proposed intelligent predictor is based on the NFS theory. Then, the FBSA clustering method is specified. The FBSA is used to automatically determine the optimal number of fuzzy If-Then rules for the intelligent predictor. Thirdly, the proposed PSO-RLSE-PSO learning method is explained, which is used to fine-tune the free parameters of the proposed predictor for accurate prediction.

2.1 Fuzzy System

To design a fuzzy intelligent predictor with multiple inputs and single output, the first-order Takagi-Sugeno (T-S) fuzzy model is used to develop a systematic approach for fuzzy rules [15]. In the first-order T-S fuzzy model with K fuzzy rules, a fuzzy rule, whose consequent part is a linear combination of crisp inputs, is given as follows.

$$\text{Rule } i\text{: IF } x_1 = {}^{i}s_1\big(h_1(t)\big) \text{ and } x_2 = {}^{i}s_2\big(h_2(t)\big) \cdots \text{and } x_M = {}^{i}s_M\big(h_M(t)\big)$$
$$\text{THEN } {}^{i}y(t) = {}^{i}a_0 + {}^{i}a_1 h_1(t) + \cdots + {}^{i}a_M h_M(t) \tag{1}$$

where the index i stands for the numeral order of the fuzzy rules from 1 to K; x_j is the jth linguistic variable, whose base (numerical) variable is denoted as h_j. $\{h_1(t), h_2(t), \cdots, h_M(t)\}$ are the numerical inputs to the NFS at time t; ${}^{i}y(t)$ is the output of the ith rule, $\{ {}^{i}s_1, {}^{i}s_2, \cdots, {}^{i}s_M\}$ are the premise fuzzy sets of the ith fuzzy rule, and $\{ {}^{i}a_0, {}^{i}a_1, \cdots, {}^{i}a_M\}$ are the consequent parameters of the ith fuzzy rule. The output of fuzzy system can be expressed as follows.

$$y^*(t) = \sum_{i=1}^{K} {}^{i}\gamma(t)\ {}^{i}y(t) \tag{2}$$

where

$${}^{i}\gamma(t) = \frac{{}^{i}\beta(t)}{\sum_{i=1}^{M} {}^{i}\beta(t)} \tag{3}$$

$${}^{i}\beta(t) = \prod_{j=1}^{M} {}^{i}\mu_j\left(h_j(t)\right) \tag{4}$$

We use Gaussian membership function in the study, defined as follows.

$$\text{gaussian}\big(h, \bar{h}, \sigma\big) = \exp\left(\frac{\big(h - \bar{h}\big)^2}{2\sigma^2}\right) \tag{5}$$

where $\bar{h}$ and σ are the mean and spread. The set of all the means and spreads are called the precise part parameters of the fuzzy rules. For the S&P 500 time series forecasting, we assume the training data for the NFS-based intelligent predictor is given as $\{X(i), i = 1, 2, \cdots n\}$. The prediction error is defined as

$$e(i) = X(i) - \hat{X}(i) \tag{6}$$

for $i = 1, 2, \cdots n$, where $X(i)$ is the observed data and $\hat{X}(i)$ is the forecast. For n prediction errors, the root mean square error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} e(i)^2} \tag{7}$$

Step 1: Set C_{min} and C_{max}.
Step 2: Initialize C_{min} cluster centers (V).
Step 3: For $c = C_{min}$ to C_{max}
  Step 3.1: Apply the basic FCM algorithm to update the membership matrix (U) and the cluster centers (V).
  Step 3.2: Test for convergence; if not converged, go to **Step 3.1**.
  Step 3.3: Compute a validity value $V_d(c)$.
  Step 3.4: Compute a score $S(i)$ for each cluster; split the worst cluster.
Step 4: Compute c_f such that the cluster validity index $V_d(c_f)$ is optimal.

Fig. 1. Implementation procedure of the FBSA, where c_f is the optimal number of clusters. Each cluster located in the input space of the NFS-based intelligent predictor represents the premise part of a T-S fuzzy rule. Thus, c_f clusters can create the same amount of T-S fuzzy If-Then rule for the predictor.

The RMSE can be used as a measure for prediction performance. And, it can be used as a cost function for the training of the fuzzy intelligent predictor. With the cost function, machine learning methods, such as PSO and RLSE, can be used to adapt the free parameters of the predictor. In the following, we continue to specify the self-organization development for the optimal number of fuzzy If-Then rules of the intelligent predictor, using the FBSA clustering method. Note that for each cluster determined by the FBSA a fuzzy If-Then rule can be created.

2.2 Structure Learning for the Proposed Predictor

The FCM-Based Splitting Algorithm (FBSA) proposed by Sun et al. [15] is a clustering algorithm based on Fuzzy C-Means and clustering validity. The general strategy adopted for the algorithm is that at each step of the algorithm it can identify the worst cluster and split it into two clusters while keeping the others. The procedure of FBSA is described in Fig.1. To determine the worst cluster, a score function $S(i)$, associated with cluster i, is applied as follows.

$$S(i) = \frac{\sum_{k=1}^{n} \mu_{ki}}{number\_of\_data\_vectors\_in\_cluster\_i} \tag{8}$$

In general, when $S(i)$ is small, the cluster tends to contain a large number of data vectors with low membership degree. The cluster with the smallest score needs to be split into two clusters. Also, the FBSA uses a validity index to decide the optimal number of clusters, given below.

$$V_d(U,V,c) = Scat(c) + \frac{Sep(c)}{Sep(C_{max})} \tag{9}$$

where $Scat(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|}{\|\sigma(x)\|}$, its value generally decreases when c increases because the clusters become more compact. The range of $Scat(c)$ is between 0 and 1. The

separation between clusters is defined as $Sep(c) = \frac{D_{min}^2}{D_{max}^2} \sum_{i=1}^{c} \left(\sum_{j=1}^{c} \|v_i - v_j\|^2 \right)^{-1}$, where $D_{min} = min_{i \neq j} \|v_i - v_j\|$ and $D_{max} = max_{i,j} \|v_i - v_j\|$.

2.3 Parameter Learning for the Proposed Predictor

After the structure learning of the predictor using the FBSA specified in the preceding subsection, we devise a hybrid learning method, called the **PSO-RLSE-PSO method**, for the parameter learning of the predictor. The original Particle Swarm Optimization (PSO) method [16] was first proposed by Kennedy and Eberhart, and it has been widely used in solving optimization problems [17]. The concept of PSO is from the behaviors for food searching by a swarm of birds. Every particle in a PSO swarm can be regarded as a bird seeking for food. The location of food can be viewed as a solution. All the particles form a population called a swarm. In the searching space, each particle owns its position and velocity, which can be adjusted by the search experiences of its own and the swarm. A particle changes its search direction according to its best location called **Pbest** and the swarm's best location called **Gbest**. To obtain the optimal solution, particles apply the idea of fitness function (or cost function) in terms of RMSE. For each particle, the velocity and position are updated as follows.

$$\begin{aligned} \mathbf{V}_i(t+1) = \omega \times \mathbf{V}_i(t) + c_1 \times rand_1 \times \left(\mathbf{Pbest}_i(t) - \mathbf{P}_i(t) \right) \\ + c_2 \times rand_2 \times \left(\mathbf{Gbest}(t) - \mathbf{P}_i(t) \right) \end{aligned} \tag{10}$$

$$\mathbf{P}_i(t+1) = \mathbf{P}_i(t) + \mathbf{V}_i(t) \tag{11}$$

where $\mathbf{V}_i(t)$ is the velocity of the ith particle, at the tth learning iteration, $\{c_1, c_2\}$ are the parameters for PSO, ω is the inertia weight, $\{rand_1, rand_2\}$ are random numbers between 0 and 1, and $\mathbf{P}_i(t)$ is the location of the ith particle.

The method of recursive least square estimator (RLSE) is basically from the least squares estimation (LSE) [12] [18] and is very efficient for the linear regression model optimization. Given a set of training data $\{(\mathbf{u}_j, y_j), j=1,2,\ldots,n\}$, a linear regression model can be given below.

$$y_j = \sum_{i=1}^{m} \theta_i f_i(u_j) + \varepsilon_j \tag{12}$$

where u_j is the input vector to the model, $\{f_i(*), i = 1, 2, \cdots, m\}$ are known functions of u, $\{\theta_i, i = 1, 2, \cdots, m\}$ are the free parameters to be estimated, and ε_j is the model error. With the RLSE method, the solution will be obtained iteratively. At iteration k, the parameter can be calculated using the following equations [13].

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \frac{\mathbf{p}_k \mathbf{a}_{k+1} \mathbf{a}_{k+1}^T \mathbf{p}_k}{1 + \mathbf{a}_{k+1}^T \mathbf{p}_k \mathbf{a}_{k+1}} \tag{13}$$

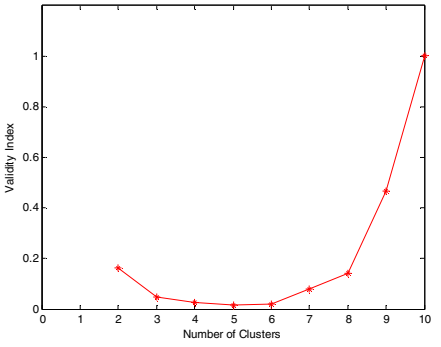


Fig. 2. The validity-index curve by the FBSA

Table 1. The validity-index result by the FBSA (Time series of daily S&P 500 index)

Amount of Clusters	Validity Index
2	0.1610
3	0.0488
4	0.0265
5	0.0175
6	0.0175
7	0.0786
8	0.1406
9	0.4687
10	1.0021

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mathbf{p}_{k+1}\mathbf{a}_{k+1}(y_{k+1} - \mathbf{a}_{k+1}^T\boldsymbol{\theta}_k) \tag{14}$$

where k is the iteration index, k=0,1,2…n-1. Before starting the RLSE, $\boldsymbol{\theta}_0$ and $\mathbf{p}_0$ are needed to be initialized. Usually, we set $\boldsymbol{\theta}_0$ to zero vector and $\mathbf{p}_0$ to $\alpha\mathbf{I}$, where α is a large number and $\mathbf{I}$ is the identity matrix.

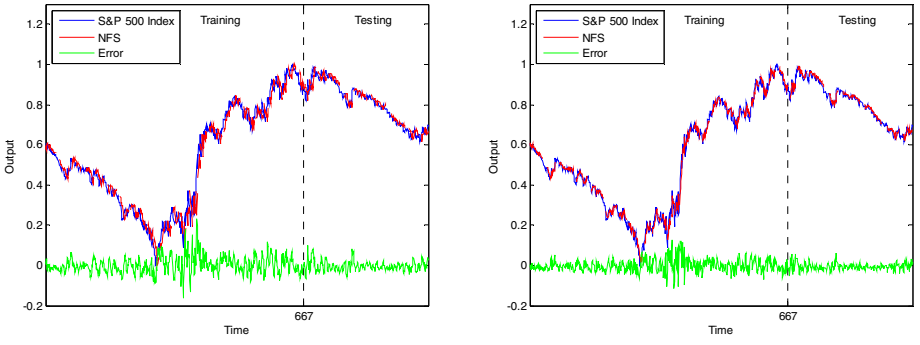
With the PSO and the RLSE specified above, we explain the proposed PSO-RLSE-PSO learning method, which is processed with two phases. In Phase 1, in hybrid way, we use PSO to adjust the premise parameters and RLSE to adjust consequent parameters. In Phase 2, the premise part parameters are fixed, and PSO is again used to adjust the consequent part parameters. In this way, the performance by the proposed predictor can be further improved. With the proposed hybrid learning method, better prediction performance has been shown in our experiments.

3 Experiments

With the proposed approach, two experiments are conducted in the section. The design of the proposed intelligent predictor is implemented by a two-stage machine learning procedure, the structure learning for stage 1 and the parameter learning for stage 2. In the structure learning stage, the FBSA method is applied to determine the optimal number of fuzzy rules for the NFS-based intelligent predictor. After this, the parameter learning stage follows to fine tune the predictor for accurate prediction performance. From the Yahoo! Finance website, the S&P 500 index data is the time series of daily closing values of from May 9th, 2006 to April 30th, 2010. According to the clustering result by the FBSA, the optimal number of fuzzy rule is five, and the validity-index curve and values are shown in Fig. 2 and Table 1. Because the data range of S&P 500 time series is wide, the data set is normalized from 0 to 1. The two third of the data set is used for training, and the rest is for testing. For prediction, we assume the stock index is related to previous records. To investigate the influence of

Table 2. The RMSE of the PSO-RLSE method and the PSO-RLSE-PSO method

Input Vector Type	(**15**)		(**16**)	
	Training	Testing	Training	Testing
PSO-RLSE Method	0.042668	0.025065	0.038689	0.022662
PSO-RLSE-PSO Method (proposed)	0.030439	0.018366	0.030105	0.017581



(a) (b)

Fig. 3. Prediction responses with the input vector type in (15) (a) by the PSO-RLSE method (b) by the PSO-RLSE-PSO method. The blue line is for the observed data, the red line is for the system responses, and the green line stands for the errors of prediction.

input order to the prediction performance, two types of input vector to the intelligent predictor are given below.

$$H(t) = [y(t-5) \quad y(t-4) \quad y(t-1)] \tag{15}$$

$$H(t) = [y(t-4) \quad y(t-1) \quad y(t-5)] \tag{16}$$

Each input variable has five fuzzy sets described by the type of Gaussian membership function. There are five fuzzy rules in the fuzzy intelligent predictor. For parameter learning, the predictor is trained by the PSO-RLSE method and the PSO-RLSE-PSO method, respectively. The settings for PSO and RLSE are given in the following. For PSO, swarm size is set to 50, initial particle positions and velocities are set randomly with Gaussian distribution, (ω, c_1, c_2) are set to $(0.5, 2, 2)$, and the max number of learning iterations is set to 500. For Phase 2 of the PSO-RLSE-PSO method, these settings are the same. For RLSE, the initial $\boldsymbol{\theta}_0$ is set to zero vector, and α is set to 10^8.

After learning, the results by the proposed approach are summarized in Table 2. For the input vector type in (15), the prediction responses by the fuzzy intelligent predictor, trained by the PSO-RLSE and the PSO-RLSE-PSO, are shown in Fig. 3, respectively. For the latter, the parameters of the fuzzy intelligent predictor after learning are given in Table 3.

Table 3. Parameters of the proposed predictor with the input vector type in (**15**) after learning by the PSO-RLSE-PSO method

Premise Part					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Center	1.55934×10^4	2.59813×10^4	6.36662×10^4	1.61281×10^3	3.10429×10^4
Spread	-5.72351×10^4	5.72053×10^4	-5.38441×10^4	2.89687×10^4	-2.07234×10^4
Center	3.89844×10^4	3.24042×10^4	-1.09144×10^4	7.94784×10^3	1.35187×10^4
Spread	-2.89523×10^4	3.40595×10^4	-4.60469×10^4	1.94799×10^4	-3.20976×10^4
Center	-5.03091×10^4	-2.95398×10^4	4.56393×10^4	-2.33065×10^2	-3.03409×10^4
Spread	5.12991×10^4	-1.69864×10^4	2.64376×10^4	-3.90619×10^4	-1.94066×10^4
Consequent Part $= a_0 + a_1h_1 + a_2h_2 + a_3h_3$					
	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
a_0	5.82179	-3.6729×10^{-1}	-5.51328	-2.6055	2.66612
a_1	-2.19147	1.07878	-9.23113×10^{-1}	-2.3991×10^{-1}	3.56367
a_2	-5.1238×10^{-1}	1.51889	3.67982	9.84084×10^{-1}	-5.6998
a_3	3.87641	5.81509×10^{-1}	9.73581×10^{-1}	-2.23418	$2..6841 \times 10^{-1}$

4 Discussion and Conclusion

In this study, a self-organizing fuzzy-based system has been used as an intelligent predictor for the S&P 500 index time series forecasting. The design of the predictor includes a two-stage procedure. The FCM-Based Splitting Algorithm (FBSA) has been used for the structure-learning stage to automatically determine the optimal number of fuzzy If-Then rules for the predictor, and the PSO-RLSE-PSO hybrid learning method has been used to fine-tune the parameters of the intelligent predictor. By the proposed approach, excellent prediction performance in terms of forecasting accuracy has been shown. The well-known PSO is an excellent method for optimization. The concept of PSO is from the simulation of social behaviors. Every particle can be regarded as a bird to seek for food which can be seen as a global solution and all the particles form a population called a swarm. In the searching space, each particle owns its position and velocity and both will be adjusted by the experience itself and the swarm's behavior. With appropriate settings, the PSO algorithm can reach the optimal solution quickly. RLSE is a widely used method in linear regression problems. It has much less computational overhead, computing time and resources, which means it needs a little time and a few resources to achieve optimization. For the purpose of fast learning, we separate the free parameters into two subsets, which are the subset of premise parameters and the subset of consequent parameters. This separation is based on the concept of divide-and-conquer to downsize the search dimension for optimization. The PSO method is applied to optimize the subset of premise parameters, and the RLSE algorithm is for the optimization of the subset of consequent parameters. In hybrid way, the two methods are combined to become the PSO-RLSE method to achieve the fast learning purpose. Because the RLSE is good at the

optimization of linear regression model only, it might not be good enough to reach the optimal solution for the problem of nonlinear time series forecasting in the study. Thus, to further investigate for optimal solution, we apply the PSO again to update the subset of consequent parameters while fixing the subset of premise parameters after the PSO-RLSE learning. This induces the proposed PSO-RLSE-PSO learning method. The experimental results in Table 2 and Fig.3 confirm that the proposed PSO-RLSE-PSO performs better than the PSO-RLSE method. The self-organizing fuzzy-based intelligent approach has shown excellent performance in accuracy for the S&P 500 index forecasting.

The main contributions of this study are threefold. First, the fuzzy intelligent prediction approach is explicable by human experience and knowledge for forecasting S&P 500, one of the most important indices of the stock markets in the US. Second, the self-organizing and swarm-based machine learning methods used in this study can avoid human interference and add simplicity and objectivity to the structure of the prediction model. Third, the proposed hybrid learning method used in the parameter learning phase can adapt the prediction model with fast learning convergence. The performance by this self-organizing fuzzy approach shows excellent prediction performance.

Acknowledgments. This research work is supported by the National Science Council, Taiwan, ROC, under the Grant contract no. NSC99-2221-E-008-088.

References

1. Chan, M.C., Wong, C.C., Lam, C.C.: Financial Time Series Forecasting by Neural Network Using Conjugate Gradient Learning Algorithm and Multiple Linear Regression Weight Initialization. Computing in Economics and Finance (61) (2000)
2. Yahoo! Finance, `http://finance.yahoo.com/q?s=^GSPC`
3. Fama, E.F., French, K.R.: The Cross-Section of Expected Stock Returns. The Journal of Finance 47, 427–465 (1992)
4. Luna, I., Soares, S., Ballini, R.: A Constructive-Fuzzy System Modeling for Time Series Forecasting. In: IEEE International Joint Conference on Neural Networks, vol. 16, pp. 2908–2913 (2007)
5. Zhao, L., Yang, Y.P.: PSO-Based Single Multiplicative Neuron Model for Time Series Prediction. Expert Systems with Applications 36, 2805–2912 (2009)
6. Rojas, I., Valenzuela, O., Rojas, F., Guillen, A., Herrera, L.J., Pomares, H., Marquez, L., Pasadas, M.: Soft-Computing Techniques and ARMA Model for Time Series Prediction. Neurocomputing 71, 519–537 (2008)
7. Wong, W.K., Xia, M., Chu, W.C.: Adaptive Neural Network Model for Time-Series Forecasting. European Journal of Operational Research 207, 807–816 (2010)
8. Reuter, U., Moller, B.: Artificial Neural Networks for Forecasting of Fuzzy Time Series. Computer-Aided Civil and Infrastructure Engineering 25, 363–374 (2010)
9. Li, C., Cheng, K.H.: Recurrent Neuro-Fuzzy Hybrid-Learning Approach to Accurate System Modeling. Fuzzy Sets and Systems 158, 194–212 (2007)
10. Li, C., Lee, C.Y.: Self-Organizing Neuro-Fuzzy System for Control of Unknown Plants. IEEE Transactions on Fuzzy Systems 11, 135–150 (2003)

11. Lee, C.C.: Fuzzy Logic in Control Systems: Fuzzy Logic Controller-Part I. IEEE Transactions on Systems, Man and Cybernetics 20, 404–418 (1990)
12. Jang, S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing: a Computational Approach to Learning and Machine Intelligence. Prentice-Hall, Upper Saddle River (1997)
13. Li, C., Hu, J.W., Chiang, T.W., Wu, T.: Computational Intelligence Hybrid Learning Approach to Time Series Forecasting. World Academy of Science, Engineering and Technology 67, 60–67 (2010)
14. Li, C., Chiang, T.W.: Complex Neuro-Fuzzy Self-learning Approach to Function Approximation. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) Intelligent Information and Database Systems. LNCS (LNAI), vol. 5991, pp. 289–299. Springer, Heidelberg (2010)
15. Sun, H.J., Wang, S.R., Jiang, Q.S.: FCM-Based Model Selection Algorithms for Determining the Number of Clusters. Pattern Recognition 37, 2027–2037 (2004)
16. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: IEEE International Joint Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995)
17. Parsopoulos, K.E., Vrahatis, M.N.: Recent Approaches to Global Optimization Problems through Particle Swarm Optimization. Natural Computing 1, 235–306 (2002)
18. Johnson, R.A.: Miller & Freund's Probability and Statistics for Engineers. Prentice-Hall, Upper Saddle River (2005)

An Efficient DCA for Spherical Separation

Hoai Minh Le[1], Hoai An Le Thi[1], Tao Pham Dinh[2], and Ngai Van Huynh[3]

[1] Laboratory of Theoretical and Applied Computer Science, LITA EA 3097
UFR MIM, University of Paul Verlaine - Metz, Ile de Saulcy, 57045 Metz, France
lethi@univ-metz.fr, lehoai@univ-metz.fr
http://lita.sciences.univ-metz.fr/lethi/
[2] Laboratory of Modelling, Optimization & Operations Research
National Institute for Applied Sciences - Rouen,
Avenue de l'Université - 76801 Saint-Etienne-du-Rouvray cedex, France
pham@insa-rouen.fr
[3] Ecole Normale Supérieur de Quinhon, Vietnam

Abstract. The binary classification problem consists in finding a separating surface minimizing an appropriate measure of the classification error. Several mathematical programming-based approaches for this problem have been proposed. The aim of spherical seperation is to find, in the input space or in the feature space, a minimal volume sphere separating set A from set B (i.e. a sphere enclosing all points of A and no points of B). The problem can be cast into the DC programming framework. Afterwards, we propose a simple DCA scheme for solving the resulting DC program in which *all computations are explicit*. Computational results show the efficiency of the proposed algorithms over the two other spherical seperation methods: FC[6] and UCM[7].

Keywords: Classification, Spherical Seperation, DC Programming, DCA.

1 Introduction

In supervised classification, the goal is to learn a function which assigns labels to arbitrary objects, given a set of already assigned independent instances. This problem is fundamental in data mining and has a vast number of applications in various domains (cancer diagnosis, document classification, text categorization,...). In this paper, we consider a specific problem that has two classes, namely, the binary classification problem. Many methods for binary classification were proposed. Support Vector Machines (SVMs) method is one of the optimization-based approaches for solving supervised machine learning problems. The basic idea of SVM is to implicitly transform data to a higher dimentional space and to find a linear binary classifier (hyperplane) without having to perform any computations in the high dimensional space. However, SVM classifiers have some limitations (for example, SVM classifiers can not preserve the distance in input space). We can directly use general nonlinear separation surfaces (spherical surface, ellipsodal surface, polyhedral surface, ...) in the input space. Few authors attempted to solve the pattern separation problem by

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 421–431, 2011.

ellipsoids ([8,9]). Ellipsoidal separator is efficient for the binary classification problems in which a class is much smaller than another. For example, in the document categorization, the set of relevant documents is usually much smaller than the set of all available documents. In ([6,7]), the authors have considered a sphere as a separation surface. In [6], the authors assume that the center of the sphere is fixed, then the problem reduces to the minimization of a convex and nonsmooth function of just one variable (optimal radius). In [7], a unconstrained moving center case is considered. The problem has been formulated as a minimisation of a nonconvexe/nondifferentiable problem. More precisely, two nonempty and disjoint finite sets of sample points in the finite dimensional space $\mathbb{R}^n$ endowed with the Euclid norm $\|\cdot\|$, say $\mathcal{A} = \{a_1, a_2, ..., a_m\}$; $\mathcal{B} = \{b_1, b_2, ..., b_k\}$ are given. The objective is to find a sphere enclosing all points of $\mathcal{A}$ and no points of $\mathcal{B}$, i.e., a sphere separating $\mathcal{A}$ from $\mathcal{B}$ (note that the role of the two sets $\mathcal{A}$ and $\mathcal{B}$ is not symmetric). If such a sphere (called $S(x, R)$-centered at x with the radius $R \in \mathbb{R}$) exists, $\mathcal{A}$ and $\mathcal{B}$ are said to be *spherically separable,* namely,

$$\|a_i - x\|^2 \leq R^2, \text{ for all } a_i \in \mathcal{A}, \tag{1}$$

$$\|b_j - x\|^2 \geq R^2, \text{ for all } b_j \in \mathcal{B}. \tag{2}$$

However, we are not able to know, in advance, whether the existence of such a sphere. We have to find a minimal error separating sphere involving a classification error function which is suitably defined. According to relations (1), (2), classification error functions can be naturally defined as follows:

$$w_\alpha(x, R) := \sum_{i=1}^{k} \max\{0, \|a_i - x\|^2 - R^2\}^\alpha + \sum_{j=1}^{m} \max\{0, -\|b_j - x\|^2 + R^2\}^\alpha, \tag{3}$$

where $\alpha \geq 1$ is a given real number. One aims to find a minimum radius sphere separating $\mathcal{A}$ from $\mathcal{B}$. By taking into account the definition of the classification error function w_α, we have to deal with the following nonconvex optimization problem:

$$(P_\alpha) \qquad \begin{cases} \min\limits_{x,R} f_\alpha(x, R) := R^{2\alpha} + C \sum_{i=1}^{k} \max\{0, \|a_i - x\|^2 - R^2\}^\alpha \\ \qquad\qquad + C \sum_{j=1}^{m} \max\{0, -\|b_j - x\|^2 + R^2\}^\alpha, \end{cases} \tag{4}$$

where the positive constant C states the relative importance of the two objectives, the radius and the classification error. Note that with $\alpha = 1$ the objective function is nonsmooth, but it is a smooth problem for $\alpha > 1$. In [7], the authors have considered the case $\alpha = 1$ and proposed a method based on DC (Difference of two Convexe) Programming and DCA for solving this problem. This DCA scheme requires to solve one quadratic program at each iteration. In our work, we deal with optimization problem (4) with $\alpha = 2$. We have the following optimization problem:

$$(\mathcal{P}_2) \qquad \begin{cases} \min\limits_{x,R} f(x, R) := R^4 + C \sum_{i=1}^{k} \max\{0, \|a_i - x\|^2 - R^2\}^2 \\ \qquad\qquad + C \sum_{j=1}^{m} \max\{0, -\|b_j - x\|^2 + R^2\}^2. \end{cases} \tag{5}$$

In this work, we attempt to use DC programming and DCA (DC Algorithm) which is a robust approach for nonconvex continuous optimisation ([2], [5]) to solve the spherical seperation problem ([5]). Our motivation is based on the fact that DCA has been successfully applied to many real wold non-convex optimization problems, especially in Machine Learning for which they provide quite often a global solution and proved to be more robust and efficient than the standard methods). Generally, DCA is a continuous approach that aims to solve a DC program that takes the form:

$$\beta_p = \inf\{F(x) := G(x) - H(x) : x \in \mathbb{R}^n\} \quad (P_{dc})$$

where G, H are lower semicontinuous proper convex functions on $\mathbb{R}^n$. Such a function F is called DC function, and $G-H$, DC decomposition of F while G and H are DC components of F. The construction of DCA involves DC components G and H but not the function F itself. Hence, for a DC program, each DC decomposition corresponds to a different version of DCA. Since a DC function F has an infinite number of DC decompositions which have crucial impacts on the qualities (speed of convergence, robustness, efficiency, globality of computed solutions,...) of DCA, the search of a "good" DC decomposition is important from algorithmic point of views.

The investigation of DC programming and DCA to the spherical separation problem ([5]) requires a rigorous study for reformulating it in term of a DC program. We first reformulate ([5]) in the form of DC program. Afterwards, we propose a simple DCA scheme for solving the resulting DC program in which *all computations are explicit*. Numerical experiments on several data sets show the efficiency of the proposed method and its superiority over two others spherical seperation methods: FC[6] and UCM[7].

The paper is organized as follows. For the reader's convenience, we provide a brief introduction to DC programming and DCA in Section 2. DC programming and DCA for solving the problem are investigated in Section 3. Preliminary computational results are reported in the last section.

2 An Introduction of DC Programming and DCA

To give the reader an easy understanding of the theory of DC programming & DCA and our motivation to use them for solving Problem $(\mathcal{P}_2)$, we briefly outline these tools in this section.

DC Programming and DCA constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization. They address the problem of minimizing a function f which is a difference of convex functions on the whole space $\mathbb{R}^p$ or on a convex set $C \subset \mathbb{R}^p$. Generally speaking, a DC program takes the form

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in \mathbb{R}^p\} \quad (P_{dc}) \tag{6}$$

where g, h are lower semicontinuous proper convex functions on $\mathbb{R}^p$. Such a function f is called DC function, and $g - h$, DC decomposition of f while g and

h are DC components of f. The convex constraint $x \in C$ can be incorporated in the objective function of (P_{dc}) by using the indicator function on C denoted χ_C which is defined by $\chi_C(x) = 0$ if $x \in C$, ∞ otherwise. Let

$$g^*(y) := \sup\{\langle x, y\rangle - g(x) : x \in \mathbb{R}^p\}$$

be the conjugate function of g. Then, the following program is called the dual program of (P_{dc}):

$$\alpha_D = \inf\{h^*(y) - g^*(y) : y \in \mathbb{R}^p\}. \quad (D_{dc}) \tag{7}$$

One can prove that $\alpha = \alpha_D$, (see e.g. [1,2]) and there is the perfect symmetry between primal and dual DC programs: the dual to (D_{dc}) is exactly (P_{dc}).

DCA is based on the local optimality conditions of (P_{dc}), namely

$$\partial h(x^*) \cap \partial g(x^*) \neq \emptyset \tag{8}$$

(such a point x^* is called *critical point* of $g - h$), and

$$\emptyset \neq \partial h(x^*) \subset \partial g(x^*). \tag{9}$$

The condition (9) is necessary local optimality of (P_{dc}). It is also sufficient for many classes of DC programs. In particular it is sufficient for the next cases quite often encountered in practice:

i) In polyhedral DC programs with h being a polyhedral convex function (see [1] - [2], [4,5] and references therein). In this case, if h is differentiable at a critical point x^*, then x^* is actually a local minimizer for (P_{dc}). Since a convex function is differentiable everywhere except for a set of measure zero, one can say that a critical point x^* is almost always a local minimizer for (P_{dc}).
ii) In case of the function f is locally convex at x^* ([2]).

The idea of DCA is simple: each iteration of DCA approximates the concave part $-h$ by its affine majorization (that corresponds to taking $y^k \in \partial h(x^k)$) and minimizes the resulting convex function (that is equivalent to determining $x^{k+1} \in \partial g^*(y^k)$).

DCA scheme
Initialization: Let $x^0 \in \mathbb{R}^p$ be a best guest, $0 \leftarrow k$.
Repeat
Calculate $y^k \in \partial h(x^k)$
Calculate $x^{l+1} \in \arg\min\{g(x) - h(x^k) - \langle x - x^k, y^k\rangle : x \in \mathbb{R}^p\}$ (P_k)
$k + 1 \leftarrow k$
Until convergence of x^k.

Convergence properties of DCA and its theoretical basis can be found in [1] - [2], [4,5], for instance it is important to mention that

- DCA is a descent method (the sequences $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing) *without linesearch*;

– If the optimal value α of the problem (P_{dc}) is finite and the infinite sequences $\{x^k\}$ and $\{y^k\}$ are bounded then every limit point x^* (resp. $\widetilde{y}$) of the sequence $\{x^k\}$ (resp. $\{y^k\}$) is a critical point of $g-h$ (resp. h^*-g^*).
– DCA has a *linear convergence* for general DC programs.
– DCA has a finite convergence for polyhedral DC programs.

It is interesting to note that ([1] - [2], [4]), DCA works with the convex DC components g and h but not the DC function f itself. Moreover, a DC function f *has infinitely many DC decompositions which have crucial impacts on the qualities* (speed of convergence, robustness, efficiency, globality of computed solutions,...) of DCA. For a complete study of DC programming and DCA the reader is referred to [1], - [2], [4,5] and references therein. The solution of a nonconvex program by DCA must be composed of two stages: the search of an *appropriate* DC decomposition and that of a *good* initial point. We shall apply *all these DC enhancement features* to solve problem $(\mathcal{P}_2)$ in its equivalent DC program given in the next section.

3 Solving Spherical Seperation Problem by DCA

3.1 DC Formulation of $(\mathcal{P}_2)$

We first reformulate $(\mathcal{P}_2)$ in the form of DC program.

Lemma 1. *Given $(x_0, R_0) \in \mathbb{R}^{n+1}$. Problem (5) is then equivalent to the following constrained optimization problem*

$$\begin{cases} \min f(x,R) \\ s.t. \quad \|x-\bar{a}\|^2 \le (1+(kC)^{-1})^{1/2} f(x_0,R_0)^{1/2}, \\ \qquad R \in [0, f(x_0,R_0)^{1/4}], \end{cases} \tag{10}$$

where $\bar{a}$ denotes the barycenter of $\mathcal{A} = \{a_1, a_2, ..., a_k\}$, i.e., $\bar{a} := \frac{1}{k}\sum_{i=1}^k a_i$.

Proof. Let (x^*, R^*) be a minimizer of problem (5).
Obviously, $0 \le R^* \le f(x^*,R^*)^{1/4} \le f(x_0,R_0)^{1/4}$. We have

$$\begin{aligned} \|x^*-\bar{a}\|^2 = \left\|\tfrac{1}{k}\textstyle\sum_{i=1}^k (x^*-a_i)\right\|^2 &\le \tfrac{1}{k}\textstyle\sum_{i=1}^k \|x^*-a_i\|^2 \\ &\le \tfrac{1}{k}\textstyle\sum_{i=1}^k \max\{0, \|x^*-a_i\|^2 - R^{*2}\} + R^{*2}. \end{aligned}$$

On the other hand, by using the Cauchy-Schwarz inequality,

$$\begin{aligned} \tfrac{1}{k}\textstyle\sum_{i=1}^k \max\{0, \|x^*-a_i\|^2 - R^{*2}\} &\le \left(\tfrac{1}{k}\textstyle\sum_{i=1}^k \max\{0, \|x^*-a_i\|^2 - R^{*2}\}\right)^{1/2} \\ &\le \left(\tfrac{f(x^*,R^*)}{kC}\right)^{1/2}. \end{aligned}$$

Combining the preceding inequalities, one obtains

$$\begin{aligned} \|x^*-\bar{a}\|^2 &\le (f(x^*,R^*))^{1/2}(kC)^{-1/2} + R^{*2} \le (1+(kC)^{-1})^{1/2} f(x^*,R^*)^{1/2} \\ &\le (1+(kC)^{-1})^{1/2} f(x_0,R_0)^{1/2}. \end{aligned}$$

□

Set

$$z := R^2;\ \gamma^2 := (1+(kC)^{-1})^{1/2} f(x_0,R_0)^{1/2};\ z_0 := f(x_0,R_0)^{1/2}$$

and denote a set $D \subseteq \mathbb{R}^{n+1}$ by

$$D := \{(x,z) \in \mathbb{R}^{n+1} :\ \ \|x-\bar{a}\|^2 \le \gamma^2;\ z \in [0,z_0]\}. \tag{11}$$

Then by Lemma 1, problem (5) is equivalent to the following problem

$$(\mathcal{P}_2^{dc}) \qquad \begin{cases} \min\limits_{(x,z)\in D} F(x,z) := z^2 + C\sum_{i=1}^{k} \max\{0, \|a_i - x\|^2 - z\}^2 \\ \qquad\qquad\qquad\qquad +C\sum_{j=1}^{m} \max\{0, -\|b_j - x\|^2 + z\}^2. \end{cases} \tag{12}$$

The following lemma allows us to formulate problem $(\mathcal{P}_2^{dc})$ as a DC programming.

Lemma 2. *Let*

$$\varphi(x,z) := C\sum_{i=1}^{k} \max\{0, \|a_i - x\|^2 - z\}^2 + C\sum_{j=1}^{m} \max\{0, -\|b_j - x\|^2 + z\}^2. \tag{13}$$

There exists $\rho_0 > 0$ (will be explicitly defined) such that the function

$$h(x,z) := \frac{\rho_0}{2}(\|x\|^2 + z^2) - \varphi(x,z),\ (x,z) \in \mathbb{R}^{n+1}$$

is a convex function on D.

Proof. Denote

$$\varphi_i(x,z) := \max\{0, \|a_i - x\|^2 - z\}^2,\ i = 1..m;$$
$$\psi(x,z) := \max\{0, -\|b_j - x\|^2 + z\}^2,\ j = 1..k.$$

Obviously, for $i = 1..k$, $j := 1..m$, one has

$$\nabla\varphi_i(x,z) = \begin{cases} 2(\|a_i - x\|^2 - z)\big(2(x-a_i), -1\big) & \text{if} \quad \|a_i - x\|^2 - z \ge 0, \\ 0 & \text{otherwise;} \end{cases} \tag{14}$$

$$\nabla\psi_j(x,z) = \begin{cases} 2(z - \|b_j - x\|^2)\big(2(b_j - x), 1\big) & \text{if} \quad z - \|b_j - x\|^2 \ge 0, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Let us show that for each $i = 1..k$, $\nabla\varphi_i$ is Lipschitzian on D. Given $(u,s),(v,t) \in D$. If $\|u-a_i\|^2 - s \ge 0$ and $\|v-a_i\|^2 - t \ge 0$ then, by using the triangle inequality, one has

$$\begin{aligned} &\|\nabla\varphi_i(u,s) - \nabla\varphi_i(v,t)\| \\ &= 2\|(\|u-a_i\|^2 - s)\big(2(u-a_i), -1\big) - (\|v-a_i\|^2 - t)\big(2(v-a_i), -1\big)\| \\ &\le 4(\|a_i - u\|^2 - s)\|u - v\| + 2(2\|v-a_i\| + 1)(|\|a_i - u\|^2 - \|a_i - v\|^2| + |s-t|) \\ &\le 4(\|a_i - u\|^2 - s)\|u - v\| \\ &+2(\|v-a_i\| + 1)\max\{\|a_i - u\| + \|a_i - v\|, 1\}(\|u-v\| + |s-t|) \\ &\le \rho(\varphi_i)\|(u,s) - (v,t)\|, \end{aligned}$$

where

$$\rho(\varphi_i) := 8(\|a_i - \bar{a}\|^2 + \gamma^2) + 4(2\|a_i - \bar{a}\| + \gamma + 1)\max\{2\|a_i - \bar{a}\| + 2\gamma, 1\}. \quad (16)$$

If $\|u - a_i\|^2 - s \geq 0$ and $\|v - a_i\|^2 - t \leq 0$ then, by the continuity of the function $(x, z) \mapsto \|x - a_i\|^2 - z$ and by the convexity of D, there exists $(y, r) \in [(u, s), (v, t)] \subseteq D$ such that $\|a_i - y\|^2 - r = 0$. Therefore, by (16), one also has

$$\begin{aligned}&\|\nabla\varphi_i(u, s) - \nabla\varphi_i(v, t)\| = \|\nabla\varphi_i(u, s) - 0\| = \|\nabla\varphi_i(u, s) - \nabla\varphi_i(y, r)\|\\ &\leq \rho(\varphi_i)\|(u, s) - (y, r)\| \leq \rho(\varphi_i)\|(u, s) - (v, t)\|.\end{aligned}$$

Hence, $\nabla\varphi_i$ is Lipschitzian on D with the Lipschitz constant $\rho(\varphi_i)$ defined by (16). Similarly, $\nabla\psi$ is Lipschizian on D with the Lipschitz constant $\rho(\psi_i)$ defined by

$$\rho(\psi_i) := 4z_0 + 4(2\|b_i - \bar{a}\| + \gamma + 1)\max\{2\|b_i - \bar{a}\| + 2\gamma, 1\}. \quad (17)$$

Consequently,

$$\nabla\varphi(x, z) = C\left(\sum_{i=1}^{k}\nabla\varphi_i(x, z) + \sum_{j=1}^{m}\nabla\psi_j(x, z)\right)$$

is Lipschitzian on D with the Lipschitz constant ρ_0 defined by

$$\rho_0 := C\left(\sum_{i=1}^{m}\rho(\varphi_i) + \sum_{j=1}^{k}\rho(\psi_j)\right). \quad (18)$$

Let $h(x, z) := \frac{\rho_0}{2}(\|x\|^2 + z^2) - \varphi(x, z),\ (x, z) \in R^{n+1}$. Then for all $(x, z),\ (y, t) \in D$, one has

$$\begin{aligned}\langle\nabla h(x, z) - \nabla h(y, t), (x, z) - (y, t)\rangle &= \rho_0\|(x, z) - (y, t)\|^2 -\\ &\langle\nabla\varphi(x, z) - \nabla\varphi(y, t), (x, z) - (y, t)\rangle \geq 0.\end{aligned}$$

Then, ∇h is a monotone operator on D. Consequently, h is a convex function on D and one completes the proof. □

3.2 DC Algorithm (DCA) for Solving $(\mathcal{P}_2)$

From the above lemma, we can propose a DC decomposition of $F(x, z)$ on D as follows:

$$F(x, z) = g(x, z) - h(x, z), \quad (19)$$

where

$$g(x, z) = \frac{\rho_0}{2}\|x\|^2 + (\frac{\rho_0}{2} + 1)z^2 \quad \text{and} \quad h(x, z) = \frac{\rho_0}{2}(\|x\|^2 + z^2) - \varphi(x, z). \quad (20)$$

Applying DC algorithm to solve problem $(\mathcal{P}_2^{dc})$ with this DC decomposition, we have the following DC algorithm.

Algorithm DCA

- **Initialization:** Select $(x_0, z_0) \in D$. $0 \leftarrow l$.
- **Repeat**
  - Compute $(y_l, t_l) = \nabla h(x_l, z_l)$ via (14) and (15).
  - $(x_{l+1}, z_{l+1}) = P_{B(\bar{a},\gamma)}(y_l/\rho_0) \times P_{[0,z_0]}(t_l/(\rho_0 + 2))$.
  - $l + 1 \leftarrow l$
- **Until** convergence of (x_l, z_l).

Note that *the projection of points onto balls and rectangles are explicitly computed*, then all computations in our DCA scheme are explicit.

3.3 Starting Point for DCA

Finding a good starting point is a challenge in designing the solution methods of DC programs by DCA. The search of such a point depends on the structure of the problem being considered. Generally, a good starting point for DCA must not be a *local minimizer*, because DCA is stationary from such a point. Nevertheless, we observe that from any initial point which is not a local minimizer, the objective function is decreasing rapidly during some first iterations of DCA. For this problem, we can define a starting point by:

– the barycenter of the set A: $x_0^{(1)} = \frac{1}{k}\sum_{i=1}^{k} a_i$.

– a point "far" from both the sets A and B (M is a sufficiently large positive constant): $x_0^{(2)} = \frac{1}{k}\sum_{i=1}^{k} a_i + M\left(\frac{1}{k}\sum_{i=1}^{k} a_i - \frac{1}{m}\sum_{j=1}^{m} b_j\right)$.

4 Computational Results

We have implemented the algorithm in the V.S C++ v6.0 environment and performed the experiments on a Intel Duo Core 3.06GHz, with 4Go of RAM. Our experiments are realized on 6 datasets taken from UCI Machine Learning Repository (Ionosphere, Bupa, Pima, Wisconsin Breast, Wisconsin WDBC, Sonar). The information about datasets is summarized in Table 1.

Table 1. Datasets

Dataset	Points	Dimension
Ionosphere	351	34
Bupa	345	6
Pima	768	8
Wisconsin Breast	683	9
Wiconsin WDBC	569	30
Sonar	208	60

Table 2. Comparative results between three algorithms

Dataset		FC([6])		UMC([7])		DCA	
		$x_0^{(1)}$	$x_0^{(2)}$	$x_0^{(1)}$	$x_0^{(2)}$	$x_0^{(1)}$	$x_0^{(2)}$
Ionosphere	Train set	80% - 0.21	88% - 0.12	**96% - 0.08**	**96% - 0.08**	89% - 0.10	89% - 0.10
	Test sest	69% - 0.30	74% - 0.26	73% - 0.25	70% - 0.29	72% - 0.25	**74% - 0.27**
	Time	0.24	0.06	118	119	0.4	0.4
Bupa	Train set	61% - 0.39	62% - 0.39	52% - 0.40	41% - 0.50	**69% - 0.31**	51% - 0.48
	Test sets	57% - 0.5	57% - 0.45	48% - 0.48	42% - 0.49	**57% - 0.48**	55% - 0.44
	Time	0.23	0.06	2.10	2.30	0.93	0.14
Pima	Train set	67% - 0.36	70% - 0.29	67% - 0.47	71% - 0.31	71% - 0.31	**76% - 0.27**
	Test sets	65% - 0.36	69% - 0.33	66% - 0.37	65% - 0.36	65% - 0.36	**70% - 0.32**
	Time	0.12	0.12	2.90	3.52	0.31	0.14
Wisconsin Breast	Train	96% - 0.03	96% - 0.03	**97% - 0.03**	91% - 0.07	**97% - 0.03**	95% - 0.05
	Test	95% - 0.04	96% - 0.03	**96% - 0.03**	89% - 0.09	**96% - 0.03**	94% - 0.05
	Time	0.25	0.27	12	7	0.48	0.28
Wisconsin WDBC	Train	86% - 0.14	91% - 0.08	89% - 0.13	**93% - 0.07**	89% - 0.12	**93% - 0.07**
	Test	78% - 0.25	90% - 0.10	86% - 0.16	90% - 0.10	87% - 0.16	**91% - 0.08**
	Time	0.4	0.4	250	150	0.7	0.4
Sonar	Train	82% - 0.17	89% - 0.10	95% - 0.03	93% - 0.06	70% - 0.27	**96% - 0.03**
	Test	46% - 0.50	48% - 0.49	53% - 0.43	49% - 0.46	51% - 0.44	**57% - 0.41**
	Time	0.48	0.51	22.5	1400	0.81	0.92

The tenfold cross-validation protocol was used. This protocol consists in splitting the dataset into ten equally size subsets. Nine of those subsets are used as training set and the remaining is considered as a test set.

We compare our algorithm DCA with two other ones:

- FC-Fixed Center([6]): the center of the sphere is fixed, thus the problem reduces to the minimization of a convex and nonsmooth function of just one variable (optimal radius).
- UMC([7]): the authors considered the problem (P_α) with $\alpha = 1$. A method based on DC Programming and DCA was proposed for solving this problem in which, at each iteration, one has to solve a quadratic program.

CPLEX 11.1 is used for solving quadratic program and linear program. In Table 2, we summarize the computational results obtained by each of three methods. For each dataset, we report the percentage of well classified points on training and test set and CPU time (in seconds). We also report the BER (Balanced Error Rate) as the average of the error rate on positive class examples and the error rate on negative class examples. Its definition is:

$$BER = \frac{1}{2}\left(\frac{\#\text{positive instances predicted wrong}}{\#\text{positive instances}} + \frac{\#\text{negative instances predicted wrong}}{\#\text{negative instances}}\right).$$

From the computational results we see that:

- In most of cases, **DCA** gives the best percentage of well classified points and BER.
- **DCA** is much faster than **UMC([7])**. In fact, in **UMC([7])**, we have to solve one quadratic program at each iteration while **DCA** only requires the explicit computations of $\partial h(x^k)$ and $\partial g^*(y^k)$.

5 Conclusion

We have rigorously studied the DC programming and DCA for the problem of spherical separation. The effect of DC decomposition are well exploited for obtaining a fast and robust algorithm. The results of **DCA** are interesting: they only require the projection of points onto balls/rectangles that is explicitly computed. The numerical results on several real data sets show that our algorithm is an efficient approach for spherical separation in large data sets and it is superior to other ones on both running-time and quality of solutions.

References

1. Le Thi, H.A., Pham Dinh, T.: Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. Journal of Global Optimization 11(3), 253–285 (1997)
2. Le Thi, H.A., Pham Dinh, T.: The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)

3. Le Thi, H. A., Pham Dinh, T., Van Ngai, H.: Exact penalty techniques in DC programming (submitted)
4. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to d.c. programming: Theory, Algorithms and Applications. Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday 22(1), 289–355 (1997)
5. Pham Dinh, T., Le Thi, H.A.: DC optimization algorithms for solving the trust region subproblem. SIAM J. Optimization 8, 476–505 (1998)
6. Astorino, A., Gaudioso, M.: A fixed-center spherical separation algorithm with kernel transformations for classification problems. Computational Management Science 6(3), 357–372 (2009)
7. Astorino, A., Fuduli, A., Gaudioso, M.: DC models for spherical separation. Journal of Global Optimization, 1–13 (2010), doi:10.1007/s10898-010-9558-0
8. Barnes, E.R.: An Algorithm for Separating Patterns by Ellipsoids, IBM. J. Res. Develop. 26, 759–764 (1982)
9. Astorino, A., Gaudioso, M.: Ellipsoidal Separation for Classification Problems. Optimizations Methods and Software 20, 267–276 (2005)

Solving an Inventory Routing Problem in Supply Chain by DC Programming and DCA

Quang Thuan Nguyen and Hoai An Le Thi

Laboratory of Theoretical and Applied Computer Science - LITA EA 3097
UFR MIM, University of Paul Verlaine - Metz, Ile de Saulcy, 57045 Metz, France
thuan@univ-metz.fr, lethi@univ-metz.fr

Abstract. Inventory routing problem (IRP) has received growing attention from both researchers and supply chain planners. It can be formulated as a mixed 0-1 nonlinear programming problem that is difficult to solve. We propose a new approach based on DC (Difference of Convex Functions) programming and DCA (DC Algorithm) for solving this challenging problem. Using an exact penalty technique and a decomposition technique, the original problem is transformed into an equivalent DC problem. DCA applied on the resulting problem gives the promising results.

Keywords: DC programming, DCA, Inventory routing problem.

1 Introduction

The inventory routing problem addresses the issue of product inventory polices and distribution plans in a cost effective manner. More precisely, given a distribution center, a set of sales-points with their demand rates and one/many vehicle(s), the objective of this problem is to determine a distribution plan that minimizes the total cost including the operation cost, the transportation cost, the delivery handling cost and the inventory holding cost.

Much research has investigated inventory routing problem ([1,2,3,4,5,6,15]). Federgruen and Zipkin [5] are probably the pioneers who investigated the integration of inventory management and routing problems. They modeled the problem as a mixed integer nonlinear program and proposed an approximation solution method. Golden et al. [6] studied a vehicle routing problem (VRP) with an inventory component and used a heuristic method. Chien et al. [2] considered the same problem and solved it by using a Lagrange dual ascent method. Recently, Aghezzaf et al. [1] proposed a mixed 0-1 nonlinear program and used column generation algorithm to solve it.

In this paper, we consider a single vehicle inventory routing problem (SIRP) - a special case of IRP, in which a single vehicle replenishes a set of sales-points from a single distribution center. It is formulated as a mixed 0-1 nonlinear program. Aghezzaf et al. [15] proposed an approach transforming (SIRP) into a sequence of mixed 0-1 linear problems (MILP) that are still NP-hard. Then, the MILPs

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 432–441, 2011.

are solved by MILP solvers. Clearly, this approach is not efficient. We propose a new approach based on DC programming and DCA to tackle directly the original mixed 0-1 nonlinear program.

DC programming and DCA are introduced by Pham Dinh in 1985 and have been extensively developed by Le Thi and Pham Dinh since 1994. Although DCA is a continuous approach, it has been investigated for solving efficiently large-scale quadratic and/or linear programming with binary variables (see e.g. [8,9,14] and the references therein) via exact penalty techniques. This motivates us to use DCA for the SIRP. Thanks to a new result concerning with exact penalty techniques in DC programming [10], we first reformulate the SIRP as a continuous optimization problem which is, in fact, a DC program. Then, we apply DCA to the resulting problem. Despite its local character, DCA with a good initial point quite often converges to global solutions in practice. In this work, a heuristic is used to seek an initial point for DCA. Moreover, in order to evaluate the solution quality of DCA, we also propose a way for determining a lower bound of the optimal value.

The rest of paper is organized as follows: The problem statement and its mathematical formulation are described in Section 2. Section 3 is dedicated to DC programming and DCA for solving the considered problem. The determination of lower bounds is presented in Section 4 while the numerical experiments are reported in Section 5. Section 6 gives some conclusions and perspectives.

2 Problem Statement and Its Mathematical Formulation

We consider an inventory routing problem in which we develop a cyclical distribution plan of a single product from a distribution center r to a set of sales-points S. The objective is to minimize the expected distribution and the inventory costs during the planning horizon without causing stock-outs at any sales-points.

Assume that the vehicle, with the capacity K, travels at an average speed ν. Each sales-point $i \in S$ has a demand rate d_i units per hour. The duration of trip from the sales-point $i \in S^+ = S \cup \{r\}$ to the sales-point $j \in S^+ = S \cup \{r\}$ is denoted by t_{ij} (in hours). The variables are following: T-the cycle time of the tour made by the vehicle; x_{ij}- the binary variable which is equal to 1 if the sales-point $j \in S^+$ is served immediately after the sales-point $i \in S^+$, and 0 otherwise; Q_{ij}-the quantity of product remaining in the vehicle when it travels from $i \in S^+$ to $j \in S^+$; and q_j- the quantity that is delivered to $j \in S$ by the vehicle.

The cost rate function of this model has four components: the fixed operating cost ϕ; the transportation cost $\frac{1}{T}\sum_{i\in S^+}\sum_{j\in S^+}(\delta\nu t_{ij}x_{ij})$, where δ is the travel cost per km; the delivery handling cost $\frac{\sum_{i\in S}\varphi_i}{T}$, where φ_i is the cost per delivery at the sales-point i; and the inventory holding cost $\sum_{i\in S}\frac{1}{2}\eta_i q_i$, where η is the holding cost per ton per hour at the sales-point i (suppose that the average stockage is $q_i/2$). The mathematical formulation, that is deduced from [15], can be written as follows:

$$(SIRP)\quad \min \phi + \frac{1}{T}\Big(\sum_{i\in S^+}\sum_{j\in S^+}(\delta\nu t_{ij}x_{ij}) + \sum_{i\in S}\varphi_i\Big) + \sum_{i\in S}\frac{1}{2}\eta_i q_i \tag{1a}$$

subject to:

$$\sum_{i\in S^+} x_{ij} = 1, \quad \forall j \in S, \tag{1b}$$

$$\sum_{i\in S^+} x_{ij} - \sum_{k\in S^+} x_{jk} = 0, \quad \forall j \in S^+, \tag{1c}$$

$$\sum_{i\in S^+}\sum_{j\in S^+} t_{ij}x_{ij} - T \le 0, \tag{1d}$$

$$\sum_{i\in S^+} Q_{ij} - \sum_{k\in S^+} Q_{jk} = q_j, \quad \forall j \in S, \tag{1e}$$

$$Q_{ij} \le K x_{ij}, \quad \forall i, j \in S^+, \tag{1f}$$

$$q_j \ge d_j T, \quad \forall j \in S, \tag{1g}$$

$$\sum_{j\in S} q_j \le K, \tag{1h}$$

$$x_{ij} \in \{0,1\}, \quad Q_{ij} \ge 0, \quad q_j \ge 0, \quad T \ge 0, \quad i, j \in S^+. \tag{1i}$$

Constraints (1b) guarantee that each sales-point is served one time. Constraints (1c) assure that the vehicle has to leave a sales-point served to the next sales-point or to the distribution center. Constraint (1d) indicates that the cycle time has to be greater than the total transportation time of the vehicle. Constraints (1e) are the delivered load balance. Constraints (1f) ensure that the quantity carried by the vehicle does not exceed the maximum capacity of the vehicle. Constraints (1g) state that the quantity delivered to a sales-point is greater than its demand. Constraint (1h) specifies that the quantity delivered to all sales-points is less than the vehicle capacity.

We adopt two assumptions: firstly, the time for loading/unloading the product is small in comparison to the travel time so that it is neglected; secondly, the inventory capacity of sales-points is so large that the corresponding capacity constraints can be omitted. Note that T is bounded in an interval $[T_{min}, T_{max}]$. T_{min} may be computed by the travel time of the travelling salesman tour (T_{TSP}) and T_{max} is determined from constraint (1h): $T_{max} = \frac{K}{\sum_{i\in S} d_i}$. Moreover, problem (SIRP) is a mixed 0-1 nonlinear program which is hard to solve.

3 Solution Method via DC Programming and DCA

3.1 DC Programming and DCA

In recent years, DC programming has been developed extensively, becoming an attractive topic of research in non-convex programming. A DC program is that of the form

$$\alpha := \min\Big\{f(x) := g(x) - h(x) : x \in \mathbb{R}^n\Big\}, \tag{2}$$

with g, h being lower semi-continuous proper convex functions on $\mathbb{R}^n$, and its dual is defined as

$$\min\Big\{h^*(y) - g^*(y) : y \in \mathbb{R}^n\Big\}, \tag{3}$$

where $g^*(y) := \max\{x^T y - g(x) : x \in \mathbb{R}^n\}$ is the conjugate function of g.

Based on local optimality conditions and duality in DC programming, the algorithm DCA consists in the construction of two sequences $\{x^k\}$ and $\{y^k\}$, candidates to be optimal solutions of primal and dual programs respectively, in a way such that $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing. The idea of DCA is simple: each iteration of DCA approximates the concave part $-h$ by its affine majorization (that corresponds to taking $y^k \in \partial h(x^k)$) and minimizes the resulting convex function.

Generic DCA scheme:
Initialization Let $x^0 \in \mathbb{R}^n$ be a best guess, $0 \leftarrow k$;
Repeat
- Calculate $y^k \in \partial h(x^k)$;
- Calculate $x^{k+1} \in \arg\min\big\{g(x) - h(x^k) - \langle x - x^k, y^k\rangle : x \in \mathbb{R}^n\big\} \quad (P_k)$;
- $k + 1 \leftarrow k$;

Until convergence of x^k.

The convergence properties of DCA and its theoretical basis can be found in [11,12,13].

3.2 DC Reformulation

Let U be the feasible set of problem (SIRP). By passing the constant ϕ, problem $(SIRP)$ becomes the following problem

$$\begin{aligned}(SIRP1)\quad \min f(z) := \frac{1}{T}\Big(\sum_{i\in S^+}\sum_{j\in S^+}(\delta\nu t_{ij}x_{ij}) + \sum_{i\in S}\varphi_i\Big) + \sum_{i\in S}\frac{1}{2}\eta_i q_i\\ \text{subject to}\\ z = (x, Q, q, T) \in U,\\ x \in \{0,1\}^{(n+1)^2}.\end{aligned}$$

Continuous formulation. By using an exact penalty result, we can reformulate (SIRP1) in the form of a continuous problem. Let U' be the set defined by $U' := U \cap \{[0,1]^{(n+1)^2} \times \mathbb{R}^{(n+1)^2+n+1}\}$. Let p be the finite function defined on U' by

$$p(z) \equiv p(x) = \sum_{i,j\in S^+} x_{ij}(1 - x_{ij}). \tag{5}$$

Clearly, the function p is nonnegative and concave on U' and

$$\Big\{z = (x, Q, q, T) \in U : x \in \{0,1\}^{(n+1)^2}\Big\} = \{z \in U' : p(z) \leq 0\}.$$

Hence, problem (SIRP1) can be rewritten as

$$\min\big\{f(z) : z = (x, Q, q, T) \in U' \text{ and } p(z) \leq 0\big\}. \tag{6}$$

Theorem 1. [10] *Let U' be a nonempty bounded polyhedral convex set in $\mathbb{R}^m$, f be a finite DC function on U' and p be a finite nonnegative concave function on U'. Then there exists $t_0 \geq 0$ such that for all $t > t_0$ the following problems have the same optimal value and the same solution set:*

$$(P_t) \quad \min\Big\{f(z) + tp(z) : z \in U'\Big\}; \quad (P) \quad \min\Big\{f(z) : z \in U', p(z) \leq 0\Big\}.$$

Proof. The proof can be found in [10]. □

Since the objective function $f(z) \in C^2(U')$ (the derivatives f' and f'' exist and are continuous on the compact set U'), it is a DC function on U' ([7]). Due to Theorem 1, with a sufficiently large number ξ, problem (6) is equivalent to:

$$\min\{F(z) := f(z) + \xi p(z) : z \in U'\}. \tag{7}$$

DC formulation. For applying DCA, we need a DC decomposition of $F(z)$. We use the following DC decomposition:

$$F(z) := g(z) - h(z) = \frac{\lambda}{2}\|z\|^2 - \Big(\frac{\lambda}{2}\|z\|^2 - F(z)\Big), \tag{8}$$

where λ is a positive number such that $h(z) = \frac{\lambda}{2}\|z\|^2 - F(z)$ is convex.

Since the function $-p$ is already convex, the function h is convex if $\overline{h}(z) = \frac{\lambda}{2}\|z\|^2 - f(z)$ is convex, i.e., its Hessian matrix is semi-definite positive.

For notational simplicity, we represent $z = (z_1, z_2, \ldots, z_{2(n+1)^2+n+1})$ for

$$(x_{00}, \ldots, x_{0n}, \ldots, x_{n0}, \ldots, x_{nn}, Q_{00}, \ldots, Q_{0n}, \ldots, Q_{n0}, \ldots, Q_{nn}, q_1, q_2, \ldots, q_n, T).$$

We have $\nabla^2\overline{h}(z) = \lambda I - \nabla^2 f(z)$ where $\big(\nabla^2 f(z)\big)_{ll'} = \frac{\partial f^2}{\partial z_l \partial z_{l'}}$ is computed as

$$\begin{cases} -\frac{\delta\nu t_{ij}}{T^2}, & \text{if } \begin{cases} (l = 1, ..., (n+1)^2 \text{and } l' \neq 2(n+1)^2 + n + 1) \text{ or} \\ (l \neq 2(n+1)^2 + n + 1 \text{and } l' = 1, ..., (n+1)^2 \end{cases}; \\ \frac{2}{T^3}\big(\sum\limits_{i,j\in S^+} \delta\nu t_{ij}x_{ij} + \sum\limits_{i\in S}\varphi_i\big), \text{ if } l = l' = 2(n+1)^2 + n + 1; \\ 0, \qquad \text{otherwise.} \end{cases} \tag{9}$$

We have $\|\nabla^2 f(z)\|_\infty = \sum\limits_{i,j\in S^+} \frac{\delta\nu t_{ij}}{T^2} + \frac{2}{T^3}\big(\sum\limits_{i,j\in S^+} \delta\nu t_{ij}x_{ij} + \sum\limits_{i\in S}\varphi_i\big)$. Hence,

$$\|\nabla^2 f(z)\|_\infty \leq \sum_{i,j\in S^+} \frac{\delta\nu t_{ij}}{T^2} + \frac{2}{T^3}\big(\delta\nu T + \sum_{i\in S}\varphi_i\big) \leq \frac{\delta\nu}{T_{min}^2}\big(\sum_{i,j\in S^+} t_{ij} + 2\big) + \frac{2\sum\limits_{i\in S}\varphi_i}{T_{min}^3}, \tag{10}$$

where T_{min} is a lower bound of T. So, the matrix $\nabla^2\overline{h}(z)$ is semi-definite positive when we choose λ satisfying the following condition:

$$\lambda \geq \Big(\frac{\delta\nu}{T_{min}^2}\big(\sum_{i,j\in S^+} t_{ij} + 2\big) + \frac{2\sum\limits_{i\in S}\varphi_i}{T_{min}^3}\Big). \tag{11}$$

Problem (7) can be now written in the form of DC program:

$$\min\{G(z) - h(z)\}, \tag{12}$$

where $G(z) := \chi_{U'}(z) + \frac{\lambda}{2}\|z\|^2$ is clearly a convex function with any value of λ satisfying (11) and $\chi_{U'}(z)$ is the indicator function of U'.

3.3 DCA for Solving (12)

According to the generic DCA scheme given above, DCA applied on the last problem (12) consists of computing the two sequences $\{u^k\}$ and $\{z^k\}$ such that

$$u^k \in \partial h(z^k); \quad z^{k+1} \in \arg\min\left\{G(z) - \langle z, u^k\rangle : z \in \mathbb{R}^{2(n+1)^2+n+1}\right\}.$$

Since the function h is differentiable, the sequence u_l^k is computed as

$$u_l^k = \begin{cases} \lambda z_l^k - \frac{\delta\nu t_{ij}}{T} + 2\xi z_l^k - \xi, & l = 1, ..., (n+1)^2\ ; \\ \lambda z_l^k, & l = (n+1)^2 + 1, ..., 2(n+1)^2; \\ \lambda z_l^k - \frac{\eta_i}{2}, & l = 2(n+1)^2 + 1, ..., 2(n+1)^2 + n; \\ \lambda z_l^k + \frac{1}{(z_l^k)^2}\Big(\sum\limits_{i,j\in S^+}(\delta\nu t_{ij}x_{ij}) + \sum\limits_{i\in S}\varphi_i\Big), & l = 2(n+1)^2 + n + 1. \end{cases} \tag{13}$$

z^{k+1} is an optimal solution of the following convex quadratic problem:

$$\min\left\{\frac{\lambda}{2}\|z\|^2 - \langle u^k, z\rangle : z \in U'\right\}. \tag{14}$$

Finally, DCA applied to (12) can be described as follows:

Algorithm DCA-IRP:
Step 1: Choose an initial point z^0 and a number $\epsilon > 0$. Set $k = 0$;
Step 2: Compute $u^k \in \partial h(x)$ via (13);
Step 3: Solve the convex quadratic problem (14) to obtain z^{k+1};
Step 4: **if** $(\|z^{k+1} - z^k\| \le \epsilon(\|z^k\| + 1)$ **then** stop, z^k is the computed solution,
else set $k = k + 1$ and go to Step 2;

The convergence of Algorithm DCA can be summarized in the next theorem.

Theorem 2. *(Convergence properties of the algorithm DCA)*
i) **DCA-IRP** *generates a sequence* $\{z^k\}$ *such that the sequence* $\{F(z^k)\}$ *is monotonously decreasing.*
ii) The sequence $\{z^k\}$ *converges to the point* z^* *which satisfies the necessary local optimality condition.*

Proof. (i) and (ii) are direct consequences of the convergence properties of general DC programs. See [11,12,13] for the detail. □

4 Computing a Lower Bound

Lemma 1. [7] *Let $p \geq 0, r \geq 0$. If $u(x), v(x)$ are convex and finite on $\mathbb{R}^n$ then the function $\max\{ru(x)+pv(x)-pr, su(x)+qv(x)-qs\}$ provides a convex minorant of the product $u(x)v(x)$ on the set $\{x \in \mathbb{R}^n | p \leq u(x) \leq q, r \leq v(x) \leq s\}$.*

Proof. See [7]. □

Consider the objective function of (SIRP1). We have

$$f(z) \geq \sum_{i\in S^+}\sum_{j\in S^+} \delta\nu t_{ij} z_{ij} + \sum_{i\in S} \frac{\varphi_i}{T} + \sum_{i\in S} \frac{1}{2}\eta_i q_i,$$

where z_{ij} is a convex minorant of the function $f_{ij} := x_{ij}\frac{1}{T}$. The functions x_{ij} and $\frac{1}{T}$ are clearly convex. Moreover, we have $0 \leq x_{ij} \leq 1$ and $\frac{1}{T_{max}} \leq \frac{1}{T} \leq \frac{1}{T_{min}}$. From Lemma 1, we get a convex minorant of the function $f_{ij} = x_{ij}\frac{1}{T}$:

$$z_{ij} = \max\{\frac{x_{ij}}{T_{max}}; \frac{x_{ij}}{T_{min}} + \frac{1}{T} - \frac{1}{T_{min}}\}. \tag{15}$$

Let α be the optimal value of the original problem. The optimal value α_1 of the following problem

$$(Relax1)\quad \alpha_1 = \min \sum_{i\in S^+}\sum_{j\in S^+} (\delta\nu t_{ij} z_{ij}) + \frac{\sum_{i\in S}\varphi_i}{T} + \sum_{i\in S}\frac{1}{2}\eta_i q_i \tag{16a}$$

subject to:

$$z_{ij} \geq \frac{x_{ij}}{T_{max}}, \quad \forall i,j \in S^+, \tag{16b}$$

$$z_{ij} \geq \frac{x_{ij}}{T_{min}} + \frac{1}{T} - \frac{1}{T_{min}}, \quad \forall i,j \in S^+, \tag{16c}$$

$$\text{and constraints (1b)- (1h)}, \tag{16d}$$

$$x_{ij} \in \{0,1\}, Q_{ij} \geq 0, q_j \geq 0, T \geq 0, z_{ij} \geq 0, \quad i,j \in S^+ \tag{16e}$$

is a lower bound of α.

In problem $(Relax1)$, we relax $x_{ij} \in \{0,1\}$ by $x_{ij} \in [0,1]$ and replace the term $\frac{1}{T}$ by its affine minorant. Since $\frac{1}{T}$ is convex in the interval $[T_{min}, T_{max}]$, we choose the tangent of $\frac{1}{T}$ at $T_0 \in [T_{min}, T_{max}]$ as an affine minorant of $\frac{1}{T}$. It is $-\frac{1}{T_0^2}T + \frac{2}{T_0}$.

So, we have a new relaxed problem:

$$(Relax2)\quad \alpha_2 = \min \sum_{i,j\in S^+} (\delta\nu t_{ij} z_{ij}) + \sum_{i\in S}\varphi_i(\frac{2}{T_0} - \frac{T}{T_0^2}) + \sum_{i\in S}\frac{1}{2}\eta_i q_i \tag{17a}$$

subject to:

$$z_{ij} \geq \frac{x_{ij}}{T_{max}}, \quad \forall i,j \in S^+, \tag{17b}$$

$$z_{ij} \geq \frac{x_{ij}}{T_{min}} - \frac{T}{T_0^2} + \frac{2}{T_0} - \frac{1}{T_{min}}, \quad \forall i,j \in S^+, \tag{17c}$$

$$\text{and constraints (16d)- (16e)}. \tag{17d}$$

Problem ($Relax2$) is a linear programming so it is easy to solve. Its optimal value α_2 is a lower bound of (IRP1) ($\alpha_2 \leq \alpha_1 \leq \alpha$).

5 Numerical Experiment

The time for computing the travel time of TSP tour is not small. In our code, T_{min} is determined by the length of the minimal spanning tree (Prim′s Algorithm) that is, in fact, a lower bound of minimal TSP tour. We take $T_0 = \frac{T_{min}+T_{max}}{2}$ and the initial point of DCA-IRP is chosen by a heuristic (see Appendix A). The algorithm DCA-IRP was coded in C and run on a Intel CPU 1.73Ghz of 2GB RAM. The commercial software CPLEX 9.1 is used for solving convex quadratic programming and linear programming.

The test data is generated as follows: the sales-points are randomly distributed over a square of 200 by 200 km. The distribution center is placed at the square center. The distance between any two sales-points is their Euclidean distance. The fixed delivery handing cost ϕ (10 euros) and the holding cost η (0.1 euro) are the same for all sales-points. The vehicle travels at the average speed 50 km/h. The transportation cost is 1 euro. We tested DCA-IRP with $n = 10, K = 100$; $n = 20, K = 200$ and $n = 30, K = 300$. Table 1, 2 and 3 show the results in which we used the following notations: VarBin- the number of binary variables; VarLin- the number of continuous variables; Contr- the number of constraints; UB- the value of the objective function obtained by DCA-IRP; CPU- the computing time of the algorithm; Iter- the iteration number of the algorithm; LB- the lower bound obtained by solving the relaxed problem ($Relax2$); and $Gap = \frac{UB-LB}{UB}100\%$.

Table 1. Results with $n = 10$ and $K = 100$

Instance	VarBin	VarLin	Contr	UB	CPU	Iter	LB	Gap
1	121	132	164	88.866	0.453	24	76.939	13.42
2	121	132	164	90.893	23.063	966	78.891	13.20
3	121	132	164	89.889	0.094	8	77.279	14.03
4	121	132	164	88.221	6.094	444	76.438	13.36
5	121	132	164	87.420	0.234	18	77.153	11.74
6	121	132	164	92.846	0.531	39	84.700	08.77
7	121	132	164	91.686	1.219	94	78.604	14.27
8	121	132	164	88.972	0.078	6	80.769	09.22
9	121	132	164	84.799	1.859	103	71.111	16.14
10	121	132	164	88.821	0.156	9	74.441	16.19
Average	121	132	164		3.378			13.03

From the result tables, we observe that:

- DCA-IRP always furnishes integer solutions although it works on a continuous domain.
- The algorithm is fast. With $n = 10$ (121 binary variables), it takes less than one second for 6/10 instances. With $n = 30$ (the problem has 961 binary variables), the problems are averagely solved in 40 seconds.

Table 2. Results with $n = 20$ and $K = 200$

Instance	VarBin	VarLin	Contr	UB	CPU	Iter	LB	Gap
1	441	462	564	113.506	2.156	35	99.539	12.31
2	441	462	564	121.532	2.797	44	108.393	10.81
3	441	462	564	116.688	2.032	33	104.176	10.72
4	441	462	564	124.376	4.547	55	110.054	11.52
5	441	462	564	119.869	2.843	34	105.058	12.36
6	441	462	564	109.476	4.703	49	97.913	10.56
7	441	462	564	124.611	4.187	28	113.702	08.75
8	441	462	564	121.762	5.406	88	110.508	09.24
9	441	462	564	117.138	5.062	79	102.551	12.45
10	441	462	564	110.562	2.141	31	99.677	09.85
Average	441	462	564		3.587			10.86

Table 3. Results with $n = 30$ and $K = 300$

Instance	VarBin	VarLin	Contr	UB	CPU	Iter	LB	Gap
1	961	992	1084	135.613	38.047	222	116.722	13.93
2	961	992	1084	135.038	44.672	240	119.570	11.45
3	961	992	1084	136.940	69.859	358	123.642	09.71
4	961	992	1084	137.389	23.985	127	117.343	14.59
5	961	992	1084	142.434	31.313	169	124.221	12.79
6	961	992	1084	143.933	28.578	133	126.003	12.46
7	961	992	1084	132.314	19.125	124	114.258	13.65
8	961	992	1084	139.348	60.560	420	120.501	13.52
9	961	992	1084	136.612	36.750	225	121.671	10.94
10	961	992	1084	139.485	31.734	179	120.269	13.78
Average	961	992	1084		38.462			12.68

- The quality of the solution obtained by DCA-IRP is good. The goodness is proved by the comparison with the lower bound. *Gap* is about 10%-13%.

6 Conclusion

In this paper, a single vehicle inventory routing problem is considered. We propose a new approach based on DC programming and DCA to solve it. The numerical results show that DCA is a fast and scalable algorithm for finding a good approximation solution: it realizes well the trade-off between efficiency and accuracy. It is so interesting to combine DCA with global approaches such as Branch and Bound for globally solving this difficult problem. This work is in progress.

References

1. Aghezzaf, E.H., Raa, B., Landeghem, H.V.: Modeling inventory routing problems in supply chains of high consumption products. European Journal of Operational Research 169, 1048–1063 (2006)

2. Chien, T.W., Balakrishnan, A., Wong, R.T.: An integrated inventory allocation and vehicle routing problem. Transportation Science 23, 67–76 (1989)
3. Dror, M., Ball, M., Golden, B.: A computational comparison of algorithms for the inventory routing problem. Annals of Operations Research 4, 3–23 (1984)
4. Dror, M., Ball, M.: Inventory/routing: Reduction from an annual to a short-period problem. Naval Research Logistics 34, 891–905 (1987)
5. Federgruen, A., Zipkin, P.: A combined vehicle routing and inventory allocation problem. Operations Research 32, 1019–1037 (1984)
6. Golden, B.L., Assad, A.A., Dahl, R.: Analysis of a large scale vehicle routing problem with an inventory component. Large Scale Systems 7, 181–190 (1984)
7. Tuy, H.: Convex Analysis and Global Optimization. Kluwer Acad. Pub., Dordrecht (1998)
8. Le Thi, H.A., Pham Dinh, T.: A continuous approach for large-scale constrained quadratic zero-one programming. (In honor of Professor ELSTER, Founder of the Journal Optimization), Optimization 50(1-2), 93–120 (2001)
9. Le Thi, H.A., Nguyen, T.P., Pham Dinh, T.: A Continuous DC Programming Approach To The Strategic Supply Chain Design Problem From Qualified Partner Set. European Journal of OperationalResearch 183, 1001–1012 (2007)
10. Le Thi H. A., Pham Dinh T., Huynh V. N.: Exact Penalty Techniques in DC Programming. Technical Report, LMI, INSA-Rouen, France (July 2007)
11. Le Thi, H.A., Pham Dinh, T.: The DC (difference of convex functions) Programming and DCA revisited with DC models of real wourd nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)
12. Pham Dinh, T., Le Thi, H.A.: Convex analysis approach to DC programming: Theory, Algorithms and Applications. Acta Mathematica Vietnamica 22(1), 289–357 (1997) (dedicated to Professor Hoang Tuy on the occassion of his 70th birthday)
13. Pham Dinh, T., Le Thi, H.: DC optimization algorithms for solving the trust region subproblem. SIAM J. Optimization 8, 476–505 (1998)
14. Pham Dinh, T., Nguyen, C.N., Le Thi, H.A.: An efficient combined DCA and B&B using DC/SDP relaxation for globally solving binary quadratic programs. Journal of Global Optimization, 38 p. (January 2010) published online
15. Zhong, Y., Aghezzaf, E.H.: Analysis and solution developement of the single vehicle inventory routing problem. In: Proc. of Modelling, computation and optimization in information systems and management sciences, Metz, France, pp. 369–378 (2008)

A Starting Point for DCA

We seek a Hamiltonian cycle via all sales-points and the distribution center. The starting binary variables for DCA are fixed by the value correspondant to the Hamiltonian cycle found. The starting continuous variables for DCA are randomly chosen. The procedure for finding a Hamiltonian cycle can be described as follows:
∘ **Input:** A distribution center r and n sales-points.
∘ **Initialize:** Choose the distribution center r is the first point of the cycle and r is the current point; all n sales-points are marked "*not visited*";
∘ **Repeat until:** No sales-point "*not visited*" exists:

∗ Choose a sales-point i "*not visited*" that is the nearest to the current sales-point and take it to the cycle;

∗ The sales-point i becomes the current sales-point and is marked "*visited*";
∘ **Output:** a Hamiltonian cycle.

A Cross-Entropy Method for Value-at-Risk Constrained Optimization

Duc Manh Nguyen[1], Hoai An Le Thi[2], Tao Pham Dinh[1]

[1] LMI, INSA of Rouen, BP 8, 76131 Mont Saint Aignan, France
[2] Laboratory of Theoretical and Applied Computer Science, University of Paul Verlaine - Metz, Ile du Saulcy, 57045 Metz, France

Abstract. In this paper, we consider a portfolio optimization problem with a Value-at-Risk constraint. It is a nonconvex nonsmooth optimization problem which is very hard to solve. We propose an approach based on the Cross-Entropy (CE) method to tackle it. The numerical results show the efficiency of our approach.

Keywords: Portfolio Optimization, Risk Management, Value at Risk, Cross-Entropy.

1 Introduction

Value-at-Risk (VaR, see e.g. [11]) is an important topic for modern financial risk management, especially due to regulatory reasons in the context of Basel-II for the banking sector, as well as Solvency-II for the insurance sector.

The Value-at-Risk of a random variable X is defined as

$$VaR_\alpha = inf\{u : F_X(u) \geq \alpha\} = F_X^{-1}(\alpha), \quad 0 < \alpha < 1,$$

where F_X is the distribution function of X.

VaR was first proposed by the global financial services firm JPMorgan Chase & Co. as a measure of acceptability for a financial position with random return. If VaR_α is taken to be quantile function of the return distribution of X, then VaR_α is said to be an acceptability functional. A higher value indicates a more acceptable, i.e. better, less risky portfolio. If on the other hand X represents the random losses, then $VaR_{1-\alpha}$ is a risk functional. High values indicate higher risk and thereby worse portfolios. See [16] for an in-depth discussion of acceptability and risk functionals. In this paper X will represent anticipated (random) returns, and therefore VaR_α is considered as an acceptability functional.

However VaR_α -being the quantile of the return distribution -has the undesirable property namely, it is nonconcave function. The non-concavity of the quantile function has two major drawbacks, one being of practical and the other of technical nature. Therefore, maximizing Value-at-Risk or minimizing a convex function under a Value-at-Risk constraint leads to a non-convex optimization problem, which consequently is hard to solve.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 442–451, 2011.

In this paper, we consider the following non-convex portfolio optimization problem for n assets:

$$\begin{cases} \max\ \mathbb{E}(x^T\xi) \\ \text{s.t.}\ \sum\limits_{i=1}^{n} x_i = 1, \\ \quad x_i \geq 0, 1 \leq i \leq n, \\ \quad VaR_\alpha(x^T\xi) \geq a, \end{cases} \tag{1}$$

where x_i denotes the relative weight of asset i in the portfolio, ξ_i the random return of asset i, $x^T\xi = \sum\limits_{i=1}^{n} x_i\xi_i$.

Because of the shortcomings of VaR mentioned, in practice it is often replaced by the Average Value-at-Risk ($AVaR$, also called Conditional Value-at-Risk) in optimization problems of the type (1) [1, 19, 23]. However, due to regulatory frameworks such as Basel II and Solvency II, Value-at-Risk remains to be an industry standard and is widely used in portfolio planning. Therefore, numerous approaches to solve problems of the form (1) either exactly or approximately have been proposed in the literature [4–6, 8–10, 15, 24]. In [4], Benati and Rizzi point out this problem to be hard (NP-complete in the strong sense), and propose a straightforward approach to solve (1) exactly based on a mixed-integer program formulation of the problem. Some heuristic solution schemes have been designed, such as random search with threshold acceptance [8, 9], or evolutionary computation techniques [10]. In [5] complete enumeration on the risk-return grid is used to find near optimal portfolios for the VaR portfolio optimization problem. Pang and Leyffer [15] formulate the problem as a linear program with equilibrium constraints to derive lower and upper bounds for a Branch-and-Bound solution. Cheon et al. [6] propose a solution technique for the more general class of probabilistically constrained linear programs which is based on a Branch-Reduced-Cut algorithm. Wozabal et al. [24] give a representation of the VaR as the difference of convex (D.C.) functions in the case finite scenario, and therefore propose Branch-and-Bound algorithm to find global optima of (1). Wozabal [25] introduced a DC Algorithm (DCA) to the D.C. formulation of the problem based on a penalty technique which is not still proved to be exact penalty.

In this paper, we propose an approach based on the Cross-Entropy (CE) method to tackle Problem (1). The CE method was motivated by an adaptive algorithm for estimating probabilities of rare events in complex stochastic network [20], which involves variance minimization. It was soon realized [21, 22] that a simple cross-entropy modification of [20] could be used not only for estimating probabilities of rare event but for solving difficult combinatorial optimization problems as well. This is done by translating the "deterministic" optimization problem into a related "stochastic" optimization problem and then using rare event simulation technique similar to [20]. Several recent application demonstrate the power of the CE method as a generic and practical tool for solving NP-hard problems. The difficulty in applying CE method is how to find a family of pdfs (probability density functions) on the feasible set of the optimization

problem and such that updating the parameters could be done as easily as possible. Due to the special structure of feasible set of problem (1), we will construct a "natural" family of pdfs on it such that CE method could be applied. The numerical results will demonstrate that our proposed algorithm finds a near-optimal solution efficiently.

The rest of paper is organized as follows. In section 2, we present in summarize about the CE method. In Section 3, we propose an application CE for our problem based on a family of exponential pdfs. Numerical experiments are reported in Section 4 while some conclusions and perspectives are discussed in Section 5.

2 The Cross-Entropy Method

The Cross - Entropy (CE) method [7, 12, 18, 20–22] has been developed by Rubinstein initially for evaluating rare events probabilities, for which a direct computation by usual methods would be unreliable. To use the CE method for solving a deterministic optimization problem, one must translate the problem into a stochastic one. The set of feasible solutions is then regarded as a set of events subjected to an importance density. Thus, using rare event simulation technique.

Suppose that we wish to maximize a real-valued performance function S over a set $\mathcal{X}$. Let us denote the maximum by γ^*, thus

$$\gamma^* = \max_{x \in \mathcal{X}} S(x). \tag{2}$$

The starting point in the methodology of the CE method is to associate an estimation problem with the optimization problem (2). To this end we define a collection of indicator functions $I_{\{S(x) \leq \gamma\}}$ on $\mathcal{X}$ for family of (discrete) probability density functions (pdfs) on $\mathcal{X}$, parameterized by a real-valued (vector) v.

For a certain $u \in V$, we consider the *associated stochastic problem* (ASP):

$$\begin{aligned} \ell(\gamma) &= \mathbf{P}_u(S(x) \geq \gamma) = \sum_{x \in \mathcal{X}} I_{\{S(x) \geq \gamma\}} f(x; u) \\ &= \mathbf{E}_u I_{\{S(x) \geq \gamma\}}, \end{aligned} \tag{3}$$

where $\mathbf{P}_u$ is the probability measure under which the random state $\mathcal{X}$ has the pdf $f(\cdot; u)$, and $\mathbf{E}_u$ denotes the corresponding expectation operator. The idea of CE method is to construct simultaneously a sequence of levels $\widehat{\gamma}_1, \widehat{\gamma}_2, ..., \widehat{\gamma}_T$ and parameters (vectors) $\widehat{v}_1, \widehat{v}_2, ..., \widehat{v}_T$ such that $\widehat{\gamma}_T$ is close to the optimal γ^* and $\widehat{v}_T$ is such that the corresponding density assigns high probability mass to the collection of states that give a high value. More specifically, we initialize by setting $v_0 = u$, choosing a not very small quantity θ, and then we proceed as follows:

1. **Adaptive updating of γ_t.** For a fixed v_t, let γ_t be the θ-quantile of $S(X)$ under v_{t-1}. That is, γ_t satisfies

$$\mathbf{P}_{v_{t-1}}(S(X) \geq \gamma_t) \geq \theta, \mathbf{P}_{v_{t-1}}(S(X) \leq \gamma_t) \geq 1-\theta, \tag{4}$$

where $X \sim f(\cdot; v_{t-1})$.

A simple estimator of γ_t, denote $\widehat{\gamma}_t$ can be obtained by drawing a random sample $X^1, X^2, ..., X^N$ from $f(\cdot; v_{t-1})$. Suppose that $S(X^{\sigma(1)}) \leq S(X^{\sigma(2)}) \leq ... \leq S(X^{\sigma(N)})$, where σ is a permutation of the set $\{1, ..., N\}$. Evaluating the $(1-\theta)$-quantile of $S(X)$ as

$$\widehat{\gamma}_t = S_{\lfloor (1-\theta)N \rfloor}. \tag{5}$$

2. **Adaptive updating of v_t.** For a fixed γ_t and v_{t-1}, derive v_t by minimizing the Kullback-Leibler distance, or equivalent to solving the next program

$$\max_v \mathbf{E}_{v_{t-1}} I_{\{S(X) \geq \gamma_t\}} W(X^i; u, v_{t-1}) \ln f(X; v), \tag{6}$$

where

$$W(x; u, v_{t-1}) = \frac{f(x; u)}{f(x; v_{t-1})}.$$

The stochastic counterpart of (6) is as follows: for fixed $\widehat{\gamma}_t$ and $\widehat{v}_{t-1}$ (the estimate of v_{t-1}), derive $\widehat{v}_t$ from following program

$$\max_v D(v) := \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X^i) \geq \gamma_t\}} W(X^i; u, v_{t-1}) \ln f(X^i; v). \tag{7}$$

In typical applications, the function D is concave and differentiable with respect to v, and thus the updating equation (7) is equivalent to solving the following system of equations:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{S(X^i) \geq \gamma_t\}} W(X^i; u, v_{t-1}) \nabla \ln f(X^i; v) = 0, \tag{8}$$

where the gradient is with respect to v.

Remark 1. Instead of updating the parameter v directly via the solution of (7) we use the following smoothed version

$$\widehat{v}_t = \alpha \widetilde{v}_t + (1-\alpha)\widehat{v}_{t-1}, t = 1, 2, ..., \tag{9}$$

where $\widetilde{v}_t$ is the parameter vector from the solution of (7), and α is called the smoothing parameter, with $0.7 \leq \alpha \leq 1$.

CE Algorithm for Optimization

1. Choose $\widehat{v}_0$, and $0 < \theta < 1$. Set $t = 1$.
2. Generate N samples $X^1, X^2, ..., X^N$ according to $f(\cdot; \widehat{v}_{t-1})$, and compute $(1-\theta)$-quantile $\widehat{\gamma}_t$ of S according to (5).
3. Using the same samples $X^1, X^2, ..., X^N$ to solve the stochastic programming (7). Denote the solution by $\widetilde{v}_t$.
4. Applying (9) to smooth out the vector $\widetilde{v}_t$.
5. Repeat step 2-4 until a pre-precified stopping criterion is met.

3 Application CE Method to Problem (1)

We consider the set

$$X = \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, x_i \geq 0, i = 1, 2, ..., n\}.$$

To apply CE, we will construct a family of pdfs $\{f(\cdot; v), v \in V\}$ on X and then use the penalty technique to treat the Value-at-Risk constraint of Problem (1) as follows

$$\text{if } x \in X, VaR_\alpha(x^T\xi) < a \text{ then } S(x) = 0,$$

where $S(x) = \mathbb{E}(x^T\xi)$ is objective function of (1). Here, we assume that the distribution of ξ (for instance, in the case finite scenario (a very common case in practice)) allows us to compute $S(x)$ and $VaR_\alpha(x^T\xi)$ uncomplicatedly.

The straightforward construction of a "natural" family of pdfs $\{f(\cdot; v), v \in V\}$ on X is in general difficult. Thus, instead of the set X, we consider the following set

$$\Omega = \{x = (x_1, x_2, ..., x_{n-1}) \in \mathbb{R}^{n-1} : \sum_{i=1}^{n-1} x_i \leq 1,\ x_i \geq 0, i = 1, 2, ..., n-1\}.$$

There exist a diffeomorphism P from X to Ω as follows

$$x = (x_1, x_2, ..., x_n) \in X \mapsto P(x) = (x_1, x_2, ..., x_{n-1}) \in \Omega. \tag{10}$$

Therefore, having a family of pdfs $\{f(\cdot; v), v \in V\}$ on Ω is equivalent to having a family of pdfs on X. The reason to choose Ω is that we can use a lot of the "natural" family of pdfs on it. In this paper, we choose a family of pdfs as follows. Firstly, we consider the exponential distribution on $\mathbb{R}^n_+$

$$f(x; v) = \exp(-\sum_{i=1}^{n} \frac{x_i}{v_i}) \prod_{i=1}^{n} \frac{1}{v_i}, \quad x \in \mathbb{R}^n_+,$$

where $v = (v_1, v_2, ..., v_n) \in \mathbb{R}^n_{++}$ is the parameter. Thus, we have

$$\int_{\mathbb{R}^n_+} f(x; v)dx = 1, \quad \forall v \in \mathbb{R}^n_{++}.$$

By considering the diffeomorphism H from $\Omega \times (0, +\infty)$ to $\mathbb{R}^n_+$

$$\begin{aligned} H: \quad & \Omega \times (0, +\infty) \to \mathbb{R}^n_+ \\ & (y_1, ..., y_{n-1}, t) \mapsto (x_1, x_2, ..., x_n), \\ & \begin{cases} x_i = ty_i, \quad i = 1, 2, ..., n-1, \\ x_n = t(1 - y_1 - ... - y_{n-1}), \end{cases} \end{aligned}$$

we have a transformation of variable for above integral.

The Jacobian matrix of the function H

$$J_H(y,t) = \begin{pmatrix} t & 0 & 0 & \dots & 0 & y_1 \\ 0 & t & 0 & \dots & 0 & y_2 \\ & & & \ddots & & \\ 0 & 0 & 0 & \dots & t & y_{n-1} \\ -t & -t & -t & \dots & -t & 1 - \sum_{i=1}^{n-1} y_i \end{pmatrix}.$$

It is clear to see that $\det(J_H(y,t)) = t^{n-1}$. Thus, we have

$$\begin{aligned} 1 &= \int_{\mathbb{R}_+^n} f(x;v)dx = \int_\Omega \left(\int_0^{+\infty} f((y,t);v) t^{n-1} dt \right) dy \\ &= \frac{1}{\prod_{i=1}^n v_i} \int_\Omega \left(\int_0^{+\infty} \exp\left(-t \sum_{i=1}^{n-1} \frac{y_i}{v_i} - \frac{t(1-y_1-\dots-y_{n-1})}{v_n} \right) t^{n-1} dt \right) dy \\ &= \int_\Omega \frac{1}{\prod_{i=1}^n v_i} \cdot \frac{(n-1)!}{\left(\sum_{i=1}^{n-1} \frac{y_i}{v_i} - \frac{(1-y_1-\dots-y_{n-1})}{v_n} \right)^n} dy. \end{aligned}$$

Now, we get a family of probability measures on Ω with the pdfs

$$g(x;v) = \frac{1}{\prod_{i=1}^n v_i} \cdot \frac{(n-1)!}{\left(\sum_{i=1}^{n-1} \frac{x_i}{v_i} - \frac{(1-x_1-\dots-x_{n-1})}{v_n} \right)^n}, \quad x = (x_1, \dots, x_{n-1}). \tag{11}$$

Therefore, by using the map P, we have a family of pdfs on X as follows

$$g(x;v) = \frac{1}{\prod_{i=1}^n v_i} \cdot \frac{(n-1)!}{\left(\frac{x_1}{v_1} + \frac{x_2}{v_2} + \dots + \frac{x_n}{v_n} \right)^n}, \quad x = (x_1, \dots, x_{n-1}, x_n) \in X. \tag{12}$$

Updating the parameter v_t

We have

$$\ln g(x;v) = \sum_{i=1}^{n-1} \ln i - n \ln\left(\sum_{i=1}^n \frac{x_i}{v_i} \right) - \sum_{i=1}^n \ln v_i.$$

Thus

$$\frac{\partial}{\partial v_j} \ln g(x;v) = \frac{1}{v_j^2} \left(\frac{n x_j}{\sum_{i=1}^n \frac{x_i}{v_i}} - v_j \right), \quad j = 1, 2, \dots, n.$$

In our case, we can solve the system of equation (8) to give a update the parameter v.

$$\sum_{i=1}^{N} I_{\{S(X_i)\geq\gamma\}} W(X^i;u,v_{t-1}) \left(\frac{nx_j}{\sum\limits_{i=1}^{n} \frac{x_i}{v_i}} - v_j \right) = 0, \quad j = 1,2,...,n,$$

therefore

$$v_j = \frac{\sum\limits_{i=1}^{N} I_{\{S(X_i)\geq\gamma\}} W(X^i;u,v_{t-1}).n.X_{ij}}{\sum\limits_{i=1}^{N} I_{\{S(X_i)\geq\gamma\}} W(X^i;u,v_{t-1}). \sum\limits_{i=1}^{n} \frac{X_{ij}}{v_j}}, \tag{13}$$

where

$$X^i = (X_{i1}, X_{i2}, ..., X_{in}) \in X, \quad i = 1,2,...,N.$$

We only consider $v = (v_1, v_2, ..., v_n)$ satisfying

$$v_j > 0, j = 1,2,...,n \quad \text{and} \quad \sum_{j=1}^{n} v_j = 1. \tag{14}$$

By solving the system of equations (13) with the condition (14), we have the formula to update v:

$$\widetilde{v}_j^t = \frac{\sum\limits_{i=1}^{N} I_{\{S(X_i)\geq\gamma\}} W(X^i;u,v_{t-1}) X_{ij}}{\sum\limits_{i=1}^{N} I_{\{S(X_i)\geq\gamma\}} W(X^i;u,v_{t-1})}, j = 1,2,...,n. \tag{15}$$

In practice, we can generate the samples on X by $g(\cdot;v)$ as follows:

1. Generate the sample $x = (x_1, x_2, ..., x_n) \in \mathbb{R}_+^n$ by $f(\cdot;v)$.
2. If $x_1 + x_2 + ... + x_n > 0$ then take $y = (y_1, y_2, ..., y_n) \in X$, where

$$y_i = \frac{x_i}{x_1 + x_2 + ... + x_n}, i = 1,2,...,n.$$

Our CE algorithm for solving Problem (1) can be described as follows

Step 1. Initialize $v_0 = u = (\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$, and $\theta \in (0,1)$, $VM = [\emptyset]$, $\epsilon > 0$.
Step 2. Draw N samples $X^1, X^2, ..., X^N$ according to $g(x;v_t)$. Compute $S(X^k), k = 1,2,...,N$. Sort the sequence $\{S(X^k)\}_{k=1}^N$ increasingly. Suppose that $S(X^{\sigma(1)}) \leq S(X^{\sigma(2)}) \leq ... \leq S(X^{\sigma(N)})$, where σ is a permutation of the set $\{1,2,...,N\}$. Set $H = \lfloor (1-\theta)N \rfloor$, then choose H best draws $X^{\sigma(H)}, X^{\sigma(H+1)}, ..., X^{\sigma(N)}$.
Step 3. Update v_{t+1} by the formula (15) and the smoothed updating (9).
Step 4. Set $M = \frac{1}{N-H+1} \sum\limits_{i=H}^{N} S(X^{\sigma(i)}), VM = [VM; M], \Delta_t = std(VM)$, where $std(V)$ computes the sample standard deviation of the data in V.
Step 5. Iterate step 2-5 until $\Delta_t < \epsilon$.

Here, M is "mean" of the best values $\{S(X^k)\}_{k=H}^N$, and VM in iteration k is a vector which contains M at all iterations less or equal k. In practice, in each iteration we should store the "best" value of the sequence $\{S(X^k)\}_{k=1}^N$, i.e., $S(X^{\sigma(N)})$, to get the "best" value when the algorithm stops.

4 Numerical Results

In this section we compare our CE algorithm with a global method. This global method bases on a new reformulation of Problem (1) as polyhedral concave programming by using an exact penalty technique [14]. We use the empirical distribution of two years of weekly data, i.e., $S = 104$ scenarios of the following

Table 1. Time frame: 2004-2005, weekly data

Name	Average Return	Variance	$VaR_{0.045}$	$AVaR_{0.045}$
CAC 40	1.0027144	0.000217647	0.9742	0.9694
Standard&Poors100	1.0004326	0.000179629	0.9755	0.9717
Nasdaq 100	1.0013597	0.000471718	0.9653	0.9504
FTSE 100	1.0021901	0.000151213	0.9824	0.9797
Hang Seng	1.0016546	0.000421260	0.9624	0.9561

Table 2. The performance of CE algorithm compared with global solutions

	CE method				Optimal solution		
a	Objective	$VaR_{0.045}$	Time (s)	Δ	Optimal	$VaR_{0.045}$	Error
0.9745	1.00270419	0.9745018	2.265625	7.1335E-06	1.00270427	0.9745	7.52E-08
0.975	1.00268620	0.9750009	2.390625	3.134E-07	1.00268702	0.975	8.19E-07
0.9755	1.00266935	0.9755004	2.265625	6.042E-07	1.00266978	0.9755	4.267E-07
0.976	1.00265231	0.9760023	2.3125	6.094E-07	1.00265254	0.976	2.331E-07
0.9765	1.00263104	0.9765280	2.375	1.5138E-06	1.00263307	0.9765	2.0349E-06
0.977	1.00260890	0.9770584	2.296875	1.7695E-06	1.00261218	0.977	3.2807E-06
0.9775	1.00258806	0.9775243	2.359375	1.5626E-06	1.00259129	0.9775	3.2317E-06
0.978	1.00256581	0.9780673	2.421875	2.1486E-06	1.0025704	0.978	4.5887E-06
0.9785	1.00254669	0.9785454	2.390625	2.0039E-06	1.00254951	0.9785	2.8197E-06
0.979	1.00252463	0.9790105	2.25	2.2734E-06	1.00252862	0.979	3.9859E-06
0.9795	1.00250197	0.9795217	2.234375	1.2756E-05	1.00250773	0.9795	5.7558E-06
0.98	1.00242424	0.9800061	2.328125	2.0580E-05	1.00248684	0.98	6.25981E-05
0.9805	1.00230423	0.9805012	2.265625	9.051E-07	1.00230464	0.9805	4.146E-07
0.981	1.00228115	0.9810232	2.25	3.1013E-06	1.00228322	0.981	2.0653E-06
0.9815	1.00226396	0.9815007	2.3125	1.8223E-06	1.00226662	0.9815	2.6617E-06
0.982	1.00224391	0.9820372	2.1875	2.78E-06	1.00224655	0.982	2.6376E-06
0.9825	1.00222225	0.9825275	2.34375	4.5333E-06	1.00222647	0.9825	4.216E-06
0.983	1.00167853	0.9830363	2.3125	0.22936342	1.00173056	0.983	5.20262E-05
0.9835	1.00155414	0.9835389	2.21875	0.01503709	1.00160072	0.9835	4.65773E-05
0.984	NA	NA	NA	NA	1.00146105	0.984	

5 indices: CAC 40, Standard&Poors 100, Nasdaq 100, FTSE 100, and Hang Seng. We take $\alpha = 0.045$.

The CE algorithm for this problem is written in MATLAB 2007, and is tested on a notebook with chipset Intel(R) Core(TM) Duo CPU 2.0 GHz, 3GB of RAM. The interesting range for a is between the VaR of the portfolio that consists only of the asset with the highest return (i.e. CAC 40 with $VaR_{0.045}$ of 0.9742 and expected weekly return 1.0027144) and the last feasible value of a, which is 0.984. And the a varies in 0.0005 steps.

In CE method, we take the number of samples $N = 2000$, $\theta = 0.01$, the parameter smooth $\alpha = 0.8$ and the number of iteration is limited to 20. The results are presented in Table (2). We observe from the numerical results that the CE algorithm furnished an ϵ-optimal solution with ϵ varies from 7.52E-08 to 6.26E-05 in all cases, except for one case $a = 0.984$ where because the event $[x \in X : VaR_\alpha(x^T\xi) \geq 0.984]$ is too rare. Moreover the CE is very fast: CPU time is less than 2.5 seconds.

5 Conclusion

In this paper, we have proposed a simple and efficient approach based on the CE method for solving a Value-at-Risk constrained Optimization. Although the theoretical convergence properties of the CE method are not yet fully understood, the computational results in Table (2) show the efficiency of the proposed approach. It finds a near-optimal solution in a shot time. In future works we plan to apply our CE algorithm for large scale setting and develop global approaches for this problem.

References

1. Andersson, F., Mausser, H., Rosen, D., Uryasev, S.: Credit risk optimization with conditional value-at-risk criterion. Mathematical Programming 89, 273–291 (2001)
2. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Thinking coherently. Risk 10(11), 68–71 (1997)
3. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Coherent measures of risk. Mathematical Finance 9, 203–228 (1999)
4. Benati, S., Rizzi, R.: A mixed integer linear programming formulation of the optimal mean/value-at-risk portfolio problem. European Journal of Operational Research 176, 423–434 (2007)
5. Campbell, R., Huisman, R., Koedijk, K.: Optimal portfolio selection in avalue-at-risk framework. Journal of Banking & Finance 25, 1789–1804 (2001)
6. Cheon, M.S., Ahmed, S., Al-Khayyal, F.: A branch-reduced-cut algorithm for the global optimization of probabilistically constrained linear programs. Mathematical Programming 108, 617–634 (2006)
7. Costa, A., Dafydd, O., Kroese, D.: Convergence properties of the cross-entropy method for discrete optimization. Operations Research Letters 35(5), 573–580 (2007)

8. Gilli, M., Kellezi, E.: A global optimization heuristic for portfolio choice with VaR and expected shortfall, in Computational methods in decision-making, economics and finance. Applied Optimization 74, 167–183 (2002)
9. Gilli, M., Kellezi, E., Hysi, H.: A data-driven optimization heuristic for downside risk minimization. The Journal of Risk 8, 1–18 (2006)
10. Hochreiter, R.: An evolutionary computation approach to scenario-based risk-return portfolio optimization for general risk measures. In: Giacobini, M. (ed.) EvoWorkshops 2007. LNCS, vol. 4448, pp. 199–207. Springer, Heidelberg (2007)
11. Jorion, P.: Value at Risk: The New Benchmark for Controlling Market Risk. McGraw-Hill, New York (2000)
12. Kroese, D.P., Porotsky, S., Rubinstein, R.Y.: The cross-entropy method for continuous multi-extremal optimization. Methodology and Computing In Applied Probability 8(3), 383–407 (2006)
13. Larsen, N., Mausser, H., Uryasev, S.: Algorithms for optimization of value-at-risk. In: Pardalos, P., Tsitsiringos, V. (eds.) Financial Engineering, e-Commerce and Supply Chain, pp. 129–157. Kluwer Academic Publishers, Dordrecht (2002)
14. Manh, N.D., Hoai An, L.T., Tao, P.D.: A new method for Value-at-Risk constrained Optimization using the DC programming and DCA. In: 24th Euro Conference on Operational Research, Lisbon (2010)
15. Pang, J.S., Leyffer, S.: On the global minimization of the value-at-risk. Optimization Methods and Software 19, 611–631 (2004)
16. Pflug, G.C., Romisch, W.: Modeling, Measuring and Managing Risk. World Scientific, Singapore (2007)
17. Pflug, G.C.: Some remarks on the Value-at-Risk and the Conditional Value-at-Risk. Probabilistic constrained optimization. Nonconvex Optimization and its Applications 49, 272–281 (2000)
18. de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A Tutorial on The Cross-Entropy Method. Annals of Operations Research 134, 19–67 (2005)
19. Rockafellar, R.T., Uryasev, S.: Optimization of Conditional Value-at-Risk. The Journal of Risk 2, 21–41 (2000)
20. Rubinstein, R.Y.: Optimization of cumputer simulation models with rare events. European Journal of Operation Reseach 99, 89–112 (1997)
21. Rubinstein, R.Y.: The simulated entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability 2, 127–190 (1999)
22. Rubinstein, R.Y.: Combinatorial optimization, cross-entropy, ants and rare events. In: Uryasev, S., Pardalos, P.M. (eds.) Stochastic Optimization: Algorithms and Application, pp. 304–358. Kluwer, Dordrecht (2001)
23. Uryasev, S.: Conditional Value-at-Risk: Optimization algorithms and applications. Financial Engineering News 14, 1–5 (2000)
24. Wozabal, D., Hochreiter, R., Pflug, G.: A d. c. formulation of value-at-risk constrained optimization. Tech. Rep. TR2008-01, Department of Statistics and Decision Support Systems, University of Vienna, Vienna (2008)
25. Wozabal, D.: A new method for value-at-risk constrained optimization using the difference of convex algorithm. Tech. Rep. TR2008-03, Department of Statistics and Decision Support Systems, University of Vienna, Vienna (2008)

Performance Comparison of Similarity Measurements for Database Correlation Localization Method

Juraj Machaj and Peter Brida

University of Zilina, Faculty of Electrical Engineering,
Department of Telecommunications and Multimedia,
Univerzitna 8215/1, 010 26 Zilina, Slovakia
{Juraj.Machaj,Peter.Brida}@fel.uniza.sk

Abstract. User positioning is a very important feature in user adaptive system. The user position can be estimated by various positioning methods. This paper investigates an impact of similarity measurements on localization error in deterministic database correlation method. It is also called fingerprinting. Main idea is to compare widely used Euclidean distance with other similarity measurements. Seven different similarity measurements are implemented to simulation model created in Matlab software tool. Computation complexity of each similarity measurement is investigated and impact of similarity measurements on localization error in normal and extreme conditions is shown.

Keywords: Database correlation method, similarity measurements, fingerprinting localization, indoor positioning.

1 Introduction

The number of LBS (Location Based Services) is rising very fast in last years [1]. Basic requirement for LBS is to know user location. That can be achieved by various ways in dependency on environment, e.g. GPS (Global Positioning System) in outdoor. On the other hand, there can be problem with high signal attenuation in indoor environment and dense urban areas. Alternative positioning solutions have to be used. Generally, they utilized various wireless communication platforms, e.g. cellular networks or IEEE 802.11x etc.

Positioning based on cellular networks is often used as alternative solution in urban environment. In urban environment, A-GPS can be also used to estimate location. Problem with localization is even bigger in indoor environment. Signal fluctuations are large because of multipath signal propagation. Many indoor localization algorithms and systems [2] based on Bluetooth [3], Zig-Bee [4], UWB (Ultra Wide Band) [5, 6], RFID [7] and IEEE 802.11x [8] were developed.

The most popular algorithms used in indoor environment are based on IEEE 802.11 [9 - 12]. Most of they uses signal strength information and are based on fingerprinting algorithm. The biggest advantage is that the algorithm does not need a new infrastructure. Another advantage seems to be multipath propagation resistance.

Fingerprinting algorithm can be implemented in various ways from mathematical point of view. They can be divided into deterministic and probabilistic algorithms. In

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 452–461, 2011.

our work we deal with fingerprinting based on deterministic algorithms based on nearest neighbor algorithm (NN), as well as more complicated algorithms k-nearest neighbours (KNN) and weighted k-nearest neighbours (WKNN) [13].

Most of researchers dealing with deterministic fingerprinting algorithms use Euclidean distance as similarity measurements between vector of RSSI (Received Signal Strength Information) collected in on-line phase and vectors stored in radio map. We try to compare Euclidean distance with another six similarity measurements and find which one is the best solution for deterministic fingerprinting localization.

Rest of paper is organized as follows. Section 2 describes related work on fingerprinting localization algorithms. In Section 3 different metrics used in our simulations are described. Simulation model created in Matlab software tool and simulation scenarios are presented in Section 4. In Section 5 simulation results are shown. Finally, Section 5 concludes the paper and provides directions for future work.

2 Related Work

The fingerprinting localization algorithms can be divided in two phases – off-line phase and on-line phase. In off-line phase radio map is created in area, where localization will be performed. In on-line phase position of mobile nodes is estimated.

Radio map construction starts by dividing area of interest into cells [14]. Each cell is represented by one reference point. In this point RSSI value from all transmitters in range – fingerprint is measured for certain period of time and stored in database.

Most of researchers in field of fingerprinting localization use deterministic approach of localization. In deterministic approach position of mobile node is computed as combination of radio map points, using:

$$\bar{x} = \sum_{i=1}^{M} \left(\omega_i . p_i \Big/ \sum_{j=1}^{M} \omega_j \right), \tag{1}$$

where p_i are coordinates of i-th reference point in radio map, ω_i and ω_j are weights and M is number of reference points stored in radio map.

The mathematical algorithm, which keeps the K biggest weights and sets the others to zero is called the WKNN (Weighted K-Nearest Neighbor) [8]. WKNN with all weights $\omega_i = 1$ is called the KNN (K-Nearest Neighbor) algorithm [14]. The simplest algorithm, where $K = 1$, is called the NN (Nearest Neighbor) [15].

One of possible weight computation is the inverse of distance between two RSSI vectors. Most of authors use Euclidean distance [8, 13-16], Junyang Zhou et al use Mahalanobis distance [17] and Binghao Li et al introduces generalized Minkowski distance [14].

3 Similarity Measurements

In this section similarity measurements that will be used in simulations are introduced. First three distances are from Minkowski distance family, next two distances

belongs to L1 family, also called the absolute difference [18]. All of these distances measure difference of two vectors. Last two types of measurements are based on correlation, which means that they measure similarity between two vectors.

3.1 Manhattan Distance

Manhattan distance is also known as city block distance, boxcar distance or absolute value distance [19]. It represents distance between points in a city road grid. It examines the absolute differences between coordinates of a pair of objects, or simply vectors. City Block distance is given by:

$$d_{Mij} = \sum_{k=1}^{n} \left| a_{ik} - b_{jk} \right|, \tag{2}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jk} represents k-th element of vector $\boldsymbol{B}$.

3.2 Euclidean Distance

Euclidean Distance is the most common use of distance. In most cases when people said about distance, they will refer to Euclidean distance. Euclidean distance or simply 'distance' examines the root of square differences between coordinates of a pair of objects. Euclidean distance is given by (3) and represents shortest distance between two vectors in Cartesian coordinate system.

$$d_{Eij} = \sqrt{\sum_{k=1}^{n} \left(a_{ik} - b_{jk} \right)^2}. \tag{3}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jk} represents k-th element of vector $\boldsymbol{B}$.

3.3 Minkowski Distance

Minkowski distance is the generalized metric distance, it is given by (4). When $m = 1$ it becomes city block distance and when $m = 2$, it becomes Euclidean distance. This distance can be used for both ordinal and quantitative variables.

$$d_{Wij} = \sqrt[m]{\sum_{k=1}^{n} \left(a_{ik} - b_{jk} \right)^m}. \tag{4}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jk} represents k-th element of vector $\boldsymbol{B}$ and m is root level.

3.4 Canberra Distance

Canberra distance examines the sum of series of a fraction differences between two vectors. Each term of fraction difference has value between 0 and 1. If one of coordinate is zero, the term become unity regardless the other value, thus the distance will not be affected.

$$d_{Cij} = \sum_{k=1}^{n} \frac{\left|a_{ik} - b_{jk}\right|}{\left|a_{ik}\right| + \left|b_{jk}\right|}. \tag{5}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jk} represents k-th element of vector $\boldsymbol{B}$.

Note that if both elements are zeros, we need to be defined as 0/0=0. This distance is very sensitive to a small change when both elements are near to zero.

3.5 Sorensen Distance

Sorensen distance is sometimes also called Bray Curtis distance. It is in fact a normalization method that is commonly used in many science fields. It views the space as grid similar to the city block distance. Sorensen distance is given by (6) and has a nice property that if all elements are positive, its value is between zero and one. Zero Sorensen distance represents exact similar vectors.

$$d_{Sij} = \sum_{k=1}^{n} \left|a_{ik} - b_{jk}\right| \Big/ \sum_{k=1}^{n} \left(a_{ik} + b_{jk}\right). \tag{6}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jk} represents k-th element of vector $\boldsymbol{B}$.

If both vectors have zero elements, the Sorensen distance is undefined. The normalization is done using absolute difference divided by the summation.

3.6 Angular Separation

Angular separation represents cosine angle between two vectors. It measures similarity rather than distance or dissimilarity. Thus, higher value of Angular separation indicates the two objects are similar. Angular separation is given by:

$$s_{Aij} = \sum_{k=1}^{n} a_{ik} \cdot b_{jk} \Bigg/ \left(\sum_{k=1}^{n} a_{ik}^2 \cdot \sum_{r=1}^{n} b_{jr}^2 \right)^{\frac{1}{2}}. \tag{7}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jr} represents r-th element of vector $\boldsymbol{B}$.

The value of angular separation is [-1, 1] similar to cosine. It is often called as Coefficient of Correlation.

3.7 Correlation Coefficient

Correlation coefficient is standardized angular separation by centering the vectors to its mean value. The value is between -1 and +1. Same as angular separation it measures similarity rather than distance or dissimilarity. Correlation coefficient is given by:

$$s_{Cij} = \sum_{k=1}^{n} \left(a_{ik} - \overline{a}_i\right) \cdot \left(b_{jk} - \overline{b}_j\right) \Bigg/ \left(\sum_{k=1}^{n} \left(a_{ik} - \overline{a}_i\right)^2 \cdot \sum_{r=1}^{n} \left(b_{jr} - \overline{b}_j\right)^2 \right)^{\frac{1}{2}}, \tag{8}$$

Where n is number of elements in vector, a_{ik} represents k-th element of vector $\boldsymbol{A}$ and b_{jr} represents k-th element of vector $\boldsymbol{B}$ and $\overline{a}_i$ and $\overline{b}_j$ are mean values of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

Correlation coefficient measures the strength and the direction of a linear relationship between two vectors.

4 Simulation Model

Simulation model created in Matlab software tool was used for investigation of impact of similarity measurement on localization error. Fingerprinting is based on signal strength measurements, therefore simulation model can be divided into two parts: radio channel and fingerprinting method. Three mathematical algorithms introduced in section 2 – NN, KNN and WKNN were implemented in the simulation model.

Received signal strength is modelled by two independent parts: path-loss and immediate variations of signal strength. Path-loss is based on multi-wall-and-floor model (MWF). The MWF model considers the nonlinear relationship between the cumulative penetration loss and the number of penetrated floors and walls. Total loss L_{WMF} in distance d can be computed from equation:

$$L_{MWF} = L_0 + 10n\log(d) + \sum_{i=1}^{I} \sum_{k=1}^{K_{wi}} L_{wik} + \sum_{j=1}^{J} \sum_{k=1}^{K_{fj}} L_{fjk}, \tag{9}$$

Where L_0 is path loss in distance of 1m in dB, n is power decay index, d is distance between transceiver and receiver in meters, I is number of walls types, K_{wi} is number of traversed walls of category i, L_{wik} is attenuation due to wall type i and k-th traversed wall in dB, J stands for number of floor types, K_{fj} is number of traversed walls of category j and L_{fjk} represents attenuation due to wall type i and k-th traversed wall in dB.

Immediate variations of signal strength could be caused by objects motion at observed area. These variations influence RSSI measurements and add measurement error. Behavior of the variations was derived from experimental measurements. Experimental measurements were performed on notebook Asus N series, with use of WirelessMon software. RSSI values from access point (AP) in range were measured 400 times. Achieved results are depicted in Fig. 1 (up chart).

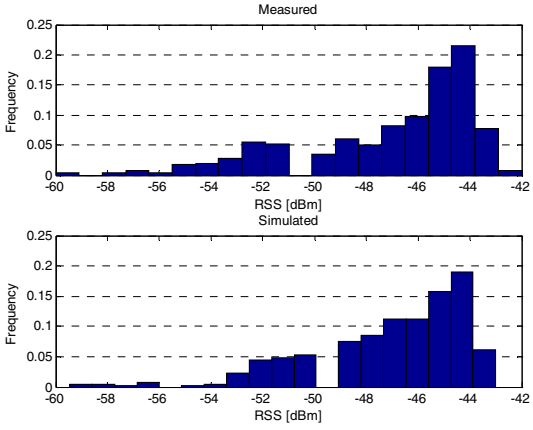


Fig. 1. Histogram of simulated and measured RSSI

According achieved measured data, immediate variations of RSSI were simulated as random variable E computed as product of two random variables with lognormal and uniform distribution respectively. Histogram of 400 simulated RSSI values is shown in Fig. 1.

Simulations of similarity measurements were done in two different scenarios. In Scenario 1 we simulated fingerprinting localization in normal conditions. This means that in on-line and off-line phase same propagation conditions were assumed. We used 6 access points to cover area of 512 square meters. Reference points were chosen in grid with 2 m distance between them. Position of mobile node is randomly chosen from all points in area. In this simulation measured RSSI values in both on-line and off-line phase were simulated with use of MWF affected with random variable E.

The problem of fading was partially eliminated by estimation of local average power in both scenarios. It is calculated as

$$\overline{RSSI} = \frac{1}{N_s}\sum_{i=1}^{N_s} RSSI_i \,, \tag{10}$$

where N_s is number of samples, in this case $N_s = 20$ was used.

In on-line phase all mathematical algorithms were used with combination of all similarity measurements. In situations when KNN and WKNN algorithms were used number of used reference point was set to 4.

In Scenario 2 we assume different environment conditions in on-line and off-line phase. Thus this scenario can be marked as extreme conditions scenario. In off-line phase, when radio map is created, fading error is simulated as random variable with uniform distribution with values from -4 dBm to 20 dBm, and in on-line phase the same distribution as in previous scenario was used. In this case, $N_s = 5$ (local average power), so fading problem is not eliminated well. All the other simulation properties were same as in Scenario 1. Simulations in both scenarios were performed with 10000 independent repetitions.

5 Simulation Results

Simulations results are introduced in this section. The results provide detailed analysis of positioning accuracy in terms of root mean square error (RMSE). The RMSE is calculated as follows:

$$RMSE = \sqrt{(x_r - x_L)^2 + (y_r - y_L)^2}\ , \tag{11}$$

where $[x_r, y_r]$ are coordinates of real (accurate) MS position and $[x_L, y_L]$ are estimated coordinates of MS computed by given mathematical algorithm.

In Fig. 2, simulation results for Scenario 1 can be seen.

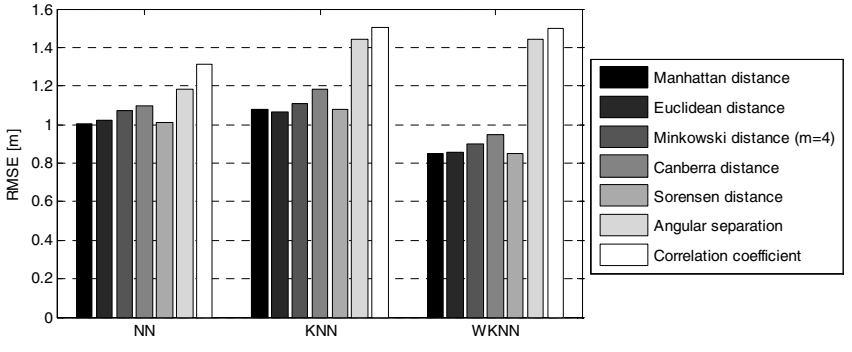


Fig. 2. Mean RMSE values for different algorithms and similarity measurements

On the basis of shown results it can be noted that particular similarity measurements have impact on RMSE regardless of mathematical algorithm. It is evident that Angular separation and Correlation coefficient metrics achieved the worst results compare with other similarity measurements. Results of remaining five metrics are almost same for individual mathematic algorithms. On the other hand, following similarity measurements Manhattan distance, Euclidean distance and Sorensen distance obtained a little bit higher accuracy. Euclidean distance performs slightly worse in combination with NN algorithm.

Difference between observed mathematical algorithms is not big from global point of view. WKNN algorithm achieved the best results (the smallest positioning error), RMSE is approximately 15 % lower. Positioning results for NN and KNN is almost same.

It is known that Manhattan and Euclidean distances are special cases of Minkowski distance. Hence, next simulation was designed to find out how Minkowski distance is affected by root coefficient *m*. Simulation results are shown in Fig. 3. It is clear that the best results are achieved in case of $m = 1.5$ for all algorithms. The greater the root coefficient *m,* the higher is the positioning error. This simulation confirms fact from previous one that WKNN is the most accurate mathematical algorithm.

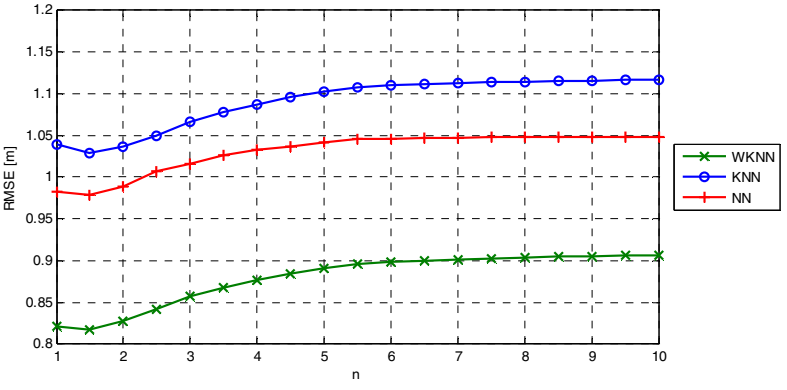


Fig. 3. RMSE values for different algorithms and root coefficient of Minkowski distance

Results of simulations in the extreme case (Scenario 2) are shown in Fig. 4. From results can be seen that best results can be achieved by same distances as in ideal case. Euclidean distance performs slightly better in combination with NN algorithm, but in more sophisticated KNN and WKNN algorithms performance of Manhattan, Euclidean, Minkowski and Sorensen distances is almost the same. Angular separation and correlation coefficient shows the worst results.

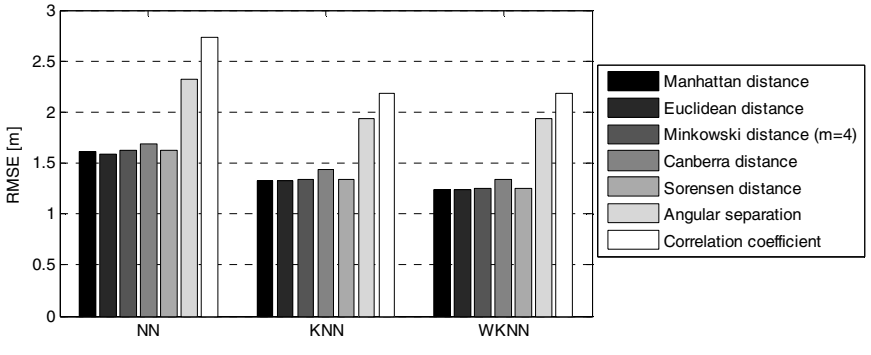


Fig. 4. Mean RMSE values for different algorithms and similarity measurements under extreme conditions

According the simulations results can be assumed that Manhattan and Sorensen distances performs well in both normal and extreme conditions, Euclidean distance performs almost same as those, difference is only with use of NN algorithm.

Last simulation results are aimed to reveal complexity of similarity measurement methods. In Fig. 5 mean computing time of each similarity measurement can be seen. From the figure it is clear that lowest complexity has Manhattan, Canberra and Sorensen distances.

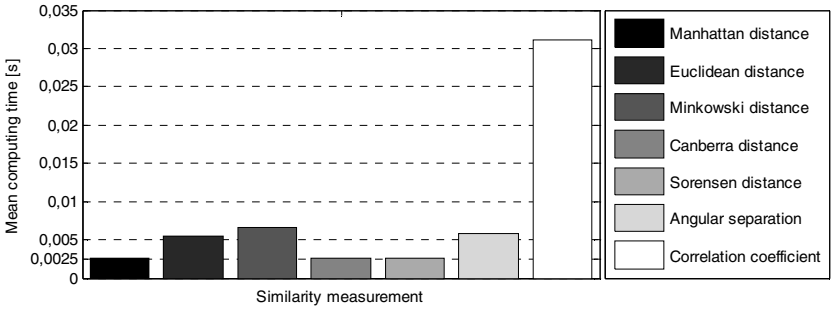


Fig. 5. Mean computing time of similarity measurements

Low complexity is very important in situations, when localization is offered for wide area with high number of users. If system must localize all of them in real time it needs to use computation methods with the lowest possible complexity.

6 Conclusion and Future Work

From results shown in this paper it is clear that Euclidean distance is not best similarity measurement for fingerprinting localization in WLAN networks. On the basis of achieved simulation results can be assumed that Manhattan and Sorensen distance performs better or same as Euclidean distance. Another advantage of Manhattan and Sorensen distances is lower complexity of computation, so these similarity measurements can be used in localization systems covering wide areas with high number of users, with better results than commonly used Euclidean distance.

For future Sorensen and Manhattan distance will be implemented into real localization system WiFiLOC, to verify results of simulation in real environment. There is also space for modification of localization algorithms and development of new mathematical algorithms to improve localization accuracy.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. LPP-0126-09 and by the Slovak VEGA grant agency, Project No. 1/0392/10 “The research of mobile nodes in wireless sensor networks”.

References

1. Mohapatra, D., Suma, S.B.: Survey of Location Based Wireless Services. In: IEEE International Conference Personal Wireless Communications 2005, pp. 358–362 (2005)
2. Hui, L., Darabi, H., Banarjee, P., Jing, L.: Survey of Wireless Indoor Positioning Techniques and Systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(6), 1067–1080 (2007), ISSN: 1094-6977

3. Chawathe, S.S.: Low-latency Indoor Localization Using Bluetooth Beacons. In: 12th International IEEE Conference Intelligent Transportation Systems, ITSC 2009, pp. 1–7 (2009), ISBN: 978-1-4244-5519-5
4. Yao, Z., Liang, D., Jiang, W., Hu, B., Fu, Y.: Implementing Indoor Positioning System via ZigBee Devices. In: Signals, Systems and Computers 2008, pp. 1867–1871 (2008)
5. Zheng, L., Dehaene, W., Gielen, G.: A 3-Tier UWB-based Indoor Localization Scheme for Ultra-low-powersensor Nodes. In: IEEE International Conference Signal Processing and Communications, ICSPC 2007, pp. 995–998 (2007), ISBN: 978-1-4244-1235-8
6. Bai, Y., Lu, X.: Research on UWB Indoor Positioning Based on TDOA Technique. In: Electronic Measurement & Instruments ICEMI 2009, pp. 167–170 (2009)
7. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC: Indoor Location Sensing Using Active RFID. Wireless Netw. 10(6), 701–710 (2004)
8. Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based User Location and Tracking System. In: INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, pp. 775–784 (2000) ISBN: 0-7803-5880-5
9. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking 2009, Article ID 802523, 8 (2009), doi:10.1155/2009/802523, ISSN: 1687-1499
10. IEEE Standard for Information Technology-telecommunications and Information Exchange Between Systems-local and Metropolitan Area Networks-specific Requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) Specifications, IEEE Standard 802.11-2007 (2007), ISBN: 978-0-7381-5656-9
11. Machaj, J., Brida, P., Tatarova, B.: Impact of the Number of Access Points in Indoor Fingerprinting Localization. In: 20th International Conference Radioelektronika, Radioelektronika 2010, pp. 83–86 (2010) ISBN 978-1-4244-6320-6
12. Krejcar, O., Frischer, R.: Detection of the Internal Defects of Material on the Basis of the Performance Spectral Density Analysis. Journal of Vibroengineering 12(4), 541–551 (2010) ISSN: 1392-8716
13. Tsung-Nan, L., Po-Chiang, L.: Performance Comparison of Indoor Positioning Techniques based on Location Fingerprinting in Wireless Networks. In: International Conference Wireless Networks, Communications and Mobile Computing 2005, vol. 2, pp. 1569–1574 (2005)
14. Li, B., Salter, J., Dempster, A. G., Rizos, C.: Indoor Positioning Techniques Based on Wireless LAN. Technical Report, School of Surveying and Spatial Information Systems, UNSW, Sydney, Australia (2006)
15. Saha, S., Chauhuri, K., Sanghi, D., Bhagwat, P.: Location Determination of a Mobile Device Using IEEE 802.11b access point signals. In: Wireless Communications and Networking, WCNC 2003, vol. 3, pp. 1987–1992 (2003), ISBN: 0-7803-7700-1
16. Yeung, W.M., Ng, J.K.: An Enhanced Wireless LAN Positioning Algorithm Based on the Fingerprint Approach. In: IEEE Region 10 Conference TENCON 2006, pp. 1–4 (2006)
17. Junyang, Z., Yeung, W.M.-C., Ng, J.K.-Y.: Enhancing Indoor Positioning Accuracy by Utilizing Signals from Both the Mobile Phone Network and the Wireless Local Area Network. In: 22nd International Conference on Advanced Information Networking and Applications, AINA 2008, pp. 138–145 (2008)
18. Cha, S.-H.: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Science 1(4) (2007)
19. Krause, E.F.: Taxicab Geometry: An Adventure in Non-Euclidean Geometry. Dover Publications, Inc., New York (1986) ISBN: 0-486-25202-7

User Perspective Adaptation Enhancement Using Autonomous Mobile Devices

Jiri Kotzian, Jaromir Konecny, and Ondrej Krejcar

VSB-Technical University of Ostrava, Department of Measurement and Control,
17. listopadu 15, 70833 Ostrava-Poruba, Czech Republic
{jiri.kotzian,jaromir.konecny,ondrej.krejcar}@vsb.cz

Abstract. The need for devices with the ability to detect toxic gases, trapped people and to multifunction has increased. Dangerous places and armed conflicts have increased the demand for remote and autonomous devices. We propose a concept of two such devices with the ability to comfortably and remotely control such devices and even with an autonomous control in remote areas inside the buildings. The localization by WiFi is used to locate a position where the GPS signal is not well presented. The ability to locate a mobile device by a wireless network is a well known possibility. The current problem is precise indoor localization where WiFi signal from the building infrastructure is not strong enough to obtain right position.

Keywords: Orientation, Navigation, Embedded system, Wireless communication.

1 Introduction

Mobile devices usually use global navigation systems like GPS, Glonass etc. for orientation in open space [1]. Navigation systems are very helpful in our everyday lives. The problem is rising in places with a high density of buildings. The precision of computing positions is too low. Inside buildings the normal navigation is usually not possible at all [2]. The reason is low signal or total signal absence. Different technologies for the navigation of mobile devices have to be used in buildings. For example: the human body uses stereovision for environment detection and orientation in cooperation with “maps” or other information (info panels, labels and indicators). Unfortunately this method is over the computation/power/space possibilities of today’s embedded systems in mobile devices. It is also possible to equip rooms of a building with a set of transmitters like GPS. But this method is very expensive and complicated. The best method would use the current data infrastructure of the building – the net of mobile Ethernet access points (WiFi, WiMax,...) [8].This method is described in the following paragraphs. A net of access points is not sufficient on its own. For obtaining the right position of mobile devices, it is necessary to equip the device with other sensors and maps. When needed we can dynamically place additional access points to achieve a higher communication range or a higher precision of position detection.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 462–471, 2011.

1.1 Motivation

Humans are well equipped with set of sensors for environment detection. However, some places are hidden to human vision, are too dangerous or too far away. As a result there is a need to equip the user with some devices which increase human perspective. This mobile device could search for people who have been trapped by an earthquake using infra camera and make audio-visual contact before the rescue come. Another example is searching for dangerous chemicals or to find criminals before authorities arrived. The simplest usage is to send or bring some item to a given position. Remote mobile robots are a common thing now, but this device should fulfill the given task using its own artificial intelligence at the end of the development. A special set of mobile devices was built at the Technical University of Ostrava for this purpose. The set is displayed in (Fig. 1).

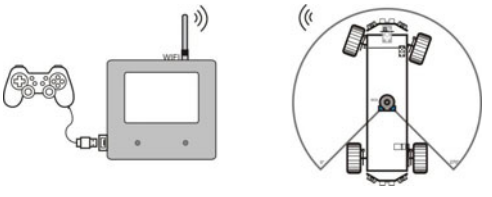


Fig. 1. The set of devices for increased human perspective

The first small handheld mobile device is the remote control and provides Human-Machine Interface to the user. The remote control is used for monitoring and controlling the second mobile device – a probe. The probe is equipped with a group of distance, pressure, temperature and other sensors for environment detection. This device is able to move by itself to a desired position sing preprogrammed scenarios and to send environment information from sensors. The two devices are connected using wireless connection. The current phase is indoor localization improvement.

2 Localization

The primary localization is needed to detect a position inside the building. Consequently, the map of the detected location is loaded into a mobile device to activate other sensors to "open the eyes" of our mobile devices. If the mobile device knows the position of the stationary device (transmitter), it also knows its own position within a range of this location provider. The typical range varies from 30 to 100 m where there is WiFi, respectively 50 m where there is BT case or 30 km for GSM. Granularity of the location can be improved by triangulation of two or more visible APs (Access Points). The mobile client currently supports the application in automatically retrieving location information from nearby WiFi, BT and GSM location providers, and in interacting with the PDPT server [3,4].

A first key step of the localization is a data collection phase. The information about the radio signals is recorded as a function of a user's location. The signal information is used to construct and validate models for signal propagation. Among other information, the signal strength (SS) is available where WiFi, BT and GSM networks are available.

To get a user position with more accuracy, the triangulation is currently used in PDPT framework. Other localization techniques like Monte Carlo localization can be used to get a better position if it is needed, but the PDPT framework provides good results only with triangulation techniques on a basic level of localization. [9]

2.1 WiFi Localization

In a real case of indoor localization by WiFi networks, several types of environments are used like open spaces, walls and mixed spaces. The Cisco APs (Cisco Aironet 1121 and 1131) are used in the test environment at the Technical University of Ostrava.

The measurements on three selected (representing three types of environment) APs of all APs have been performed to get signal strength (SS) characteristics. The characteristics were combined to get a one characteristic called "Super-Ideal characteristic". The computed equation for Super-Ideal characteristic is taken as basic equation for PDPT Core to compute the real distance from WiFi SS. The equation has a sufficient coefficient of determination $R^2 = 80$ % (R^2= ssreg/sstotal).

In the case of in-door location the damping effect of walls especially when the number of BS´s is small could hamper the positioning. However the precise positioning is not needed in all cases. When the granularity of object areas to be prebuffered into the mobile device cache is in level less than tens of meters, the localization by one or two visible BS´s is possible with high level of success. Maximal location error for static localization is 25 meters for the case the only 1 WiFi AP is in the range, 7 meters for 2 APs in range, resp. less than 5 meters in case of 3 or more WiFi APs (mentioned Cisco models) in range. This localization error can be rapidly reduced by use of dynamic localization in a sense of user movement trajectory computation. Naturally, this localization principle can be applied to other wireless technologies like Bluetooth, GSM or WiMAX.

2.2 Odometry Localization

The primary localization is made through a WiFi infrastructure. However, the accuracy of position for mobile robot is insufficient. As a result WiFi localization can be used for roughly localization – which room the device is in. Nevertheless, the WiFi signal is not well presented in all places.

Odometry was tested, due to get more precise position of mobile device. Odometry is the use of data from moving sensors to estimate change in position over time. The probe simply count route based on incremental sensors from all wheels and steer angle. The practical results show that real precision of position is from 1 to 10% with

integral character depending on surface conditions and trajectory. In cooperation with distance sensors and maps it is possible to correct inaccuracy of the method.

2.3 Magnetic Field Measuring

The magnetic field sensor could be one of sensor which is able to correct inaccuracy of previous method. The electronic compass was used for measuring azimuth of the probe. The probe was placed on 9 places with the same azimuth. The azimuth was measured on these places using electronics compass and magnetic compass. (Fig. 2) shows measuring result.

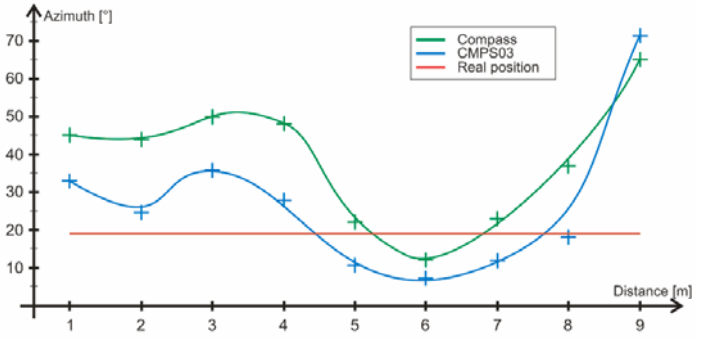


Fig. 2. Indoor Magnetic field measurement

In spite of expectations, practical measurements inside the building shows high influence of metal construction of the building and influence of electric wiring in walls. As a result we can say that this method is not possible to use to obtain more precision position.

2.4 Laser Scanner Measuring

The current work on more precise position is to use distance sensors. The probe is equipped with laser scanner, two pieso-sensors and four corner optical distance sensors. Estimation of the position within a room is done using of fingerprinting localization. For similarity measurement is used correlation between fingerprint and current measurement. Data from laser sensors is array of distances between the sensor and closest subject within given angle. The range is 270° and maximum distance is 20m. The space around the sensor can be displayed by connecting measurement distances/angle. The C# implementation of used formula [1] is shown in (Fig. 3).

$$(f * g)_k = \sum_{i=0}^{\infty} f_i \cdot g_{k+i} \tag{1}$$

```
public Measure XCorr(long[] refer, long[] test)
{
    Measure ret = new Measure();
    for (int n = 0; n < refer.Length; n++)
    {
        long k = 0;
        for (int m = 0; m < refer.Length; m++)
        {
            int i = m + n;
            if (i >= refer.Length)
                i -= refer.Length;
            k += refer[m] * test[i];
        }
        ret.Values[n] = (int)(k / 10000);
    }
    return ret;
}
```

Fig. 3. Implementation of similarity measurement

The measured data from the laboratory of embedded system are shown on (Fig.4). The room is approx. 3m x 4m. The open door can be seen as long rectangle in bottom part. Right picture shows probe rotated (angle 30.75°). The angle is computed using correlation. The correlation curve is displayed on right part of the picture. The measured angle is found as global maximum of the curve.

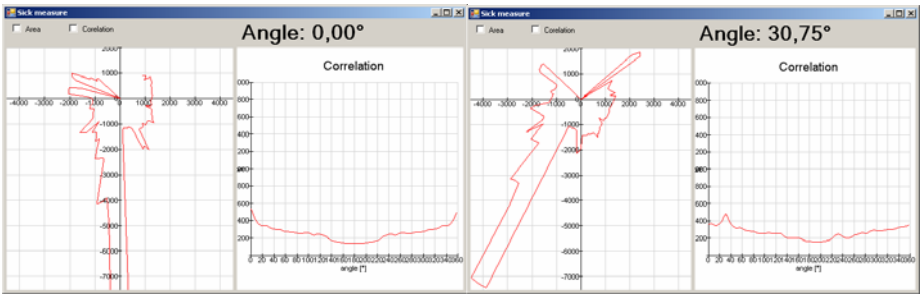


Fig. 4. An angle measurement of the probe

The measurement is influenced by the difference between fingerprint and current state of the environment. The error in measurement is increasing especially due to placing new furniture or due to movement of people. The given method was tested by movement of a man in different distance from the sensor. The measurement results are shown on (Fig. 5). Left picture represent error when human is 30cm from the sensor (minimal due to size of the probe); right picture display situation when human is 1m from the sensor. A human stands in the worst part of the area due to masking view into the corridor. The corridor has high weight in correlation in the case.

As a result we can say that this method is very suitable to obtain more precise position within given room. The method could also be used for determine if estimated position from WiFi infrastructure is right due to correlation of measured data to expected fingerprint.

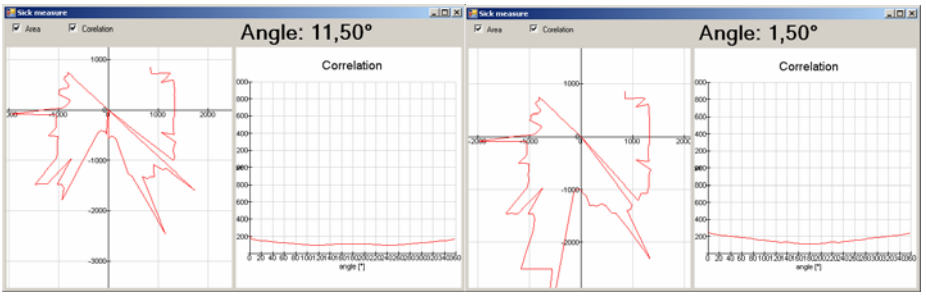


Fig. 5. An angle measurement error of the probe with disturbance 30cm and 1m from the sensor

3 Remote Control

The remote control is the first mobile device from the set. The remote control provides Human-Machine Interface. The remote control is a medium sized handheld device. Remote control provides monitor and control of the probe, display information from sensors, setting the task or the desired position in auto mode, manual control using a gamepad or joystick, maps insertion and actualization, audio-visual interface, diagnostic interface with trends and help.

The user has full control of the second mobile device (the probe) or he can give the task to the probe. The user can handle the probe using a color display and touch panel. Manual mode is also available. The probe is controlled by a gamepad in manual mode. Programmed application provides intuitive user interface with a set of buttons, value displays and screens. First picture (Fig. 6a.) displays the situation measured using distance sensors – laser, optical, and pieso. In the situation something is very close to the back of the probe. The only free space to ride is on the right side. Next three pictures (Fig.6a to Fig.7b.) illustrate other screens of the remote control. [5]

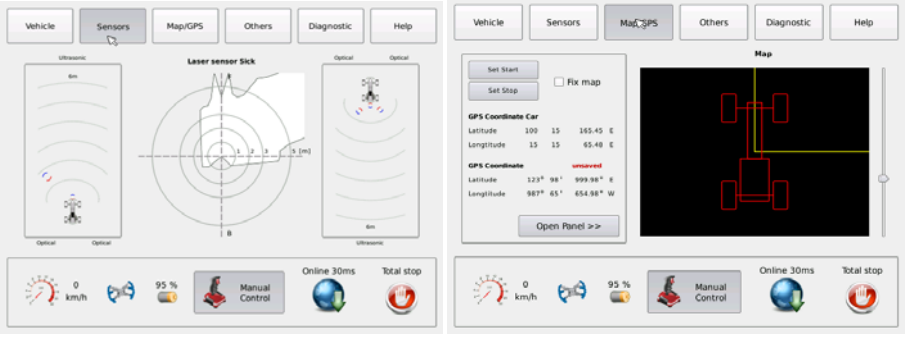


Fig. 6. Sensors screen from remote control application – laser, pieso and infra distance sensors (left figure 6a). Position of the probe and maps of remote control application (right figure 6b).

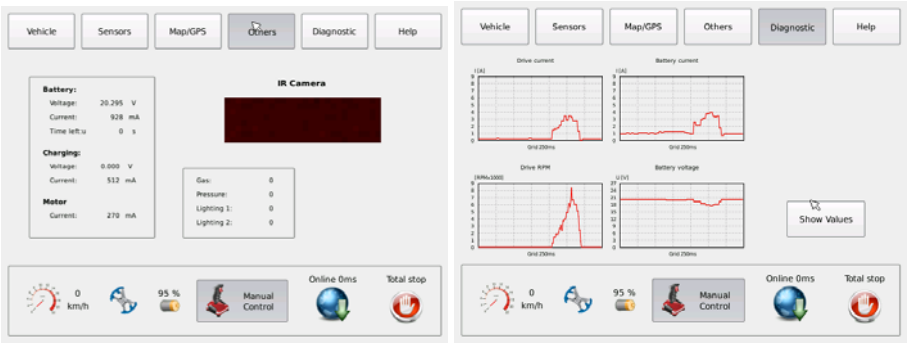


Fig. 7. Camera and diagnostic screen from remote control application (left figure 7a). Diagnostic screen from remote control application – trends or list of values (right figure 7b).

3.1 Architecture of the Remote Control

The remote control uses the iMX31LiteKit embedded controlling board based on the ARM architecture. The core of the board is the Freescale iMX31 microprocessor. The embedded board is equipped with a set of interfaces (Ethernet, serial, SPI, SD, CF etc.) The controlling application is stored on external SD card.

For debugging purposes, the device is equipped with Ethernet connection. The remote control communicates with the probe using the WiFi module Owspa311 connected via serial interface. The remote control uses a touch panel placed on the color display with a resolution of 640x480 pixels. It is possible to use a gamepad or joystick in the case of manual mode. The architecture is displayed in (Fig. 8).

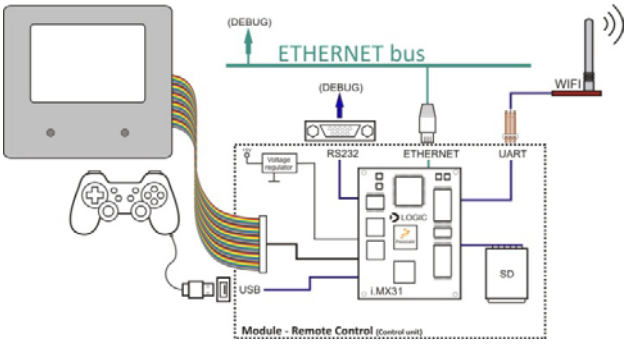


Fig. 8. Architecture of the Remote Control – display, mainboard and WiFi communication

3.2 SW Architecture of the Remote Control

The software application is based on the LinuxLink embedded linux. The application is programmed and compiled using the TimeStorm integrated development environment (IDE) including board support package (BSP) for the iMX31Litekit. The Fedora Linux host machine and the NFS (network file system) are used for developing the application. This way is quite complicated but very fast. The

application is compiled and stored on the NFS or on an SD card. After reset (power on) the Logic Loader loads the application from the SD card and starts it.

The application is developed in the Qt graphic tool – Qt Creator. In this tool it is possible to create the design, windows, buttons and main root. To create the final application it is necessary to select an external compiler and to make the application. Another way is to use the Qt designer, which only generates functions and windows. The application is then programmed and compiled in some IDE environment – the TimeStorm in the case of the iMX31.

4 Probe

The second mobile device – the probe – is based on a massive chassis with a distributed control system, motor and high capacity battery. It is approximately 85cm meter long and 60cm high without an antenna. The weight is approximately 10kg including the DC drive, the battery, the CAN based distributed control system and all sensors.

The probe includes a set of sensors for distance measuring, environment measurement and position and movement detection. For example, the Laser scanner can measure an environment up to 20meters in 270 degrees. The probe includes infra-camera, pieso and optical distance sensors, pressure, temperature, GAS sensors, 3-axis accelerometers etc. The audio-visual interface is in the preparation phase. [5,6]

4.1 Architecture of the Probe

The probe is equipped with a distributed control system with a set of embedded control and monitoring boards. The system is based on the industrial CAN bus and the CANOpen application layer.

The main control unit uses the same HW architecture like the remote control. The control unit is based on the iMX31LiteKit and it is also equipped with LinuxLink (Fig. 9).

Other control boards are based on industrial microcontrollers Freescale HCS12 with the cooperation of the FreeRTOS operating system. These boards are programmed using C programming language using the CodeWarrior IDE. The probe communicates with the remote control using the WiFi module Owspa311.

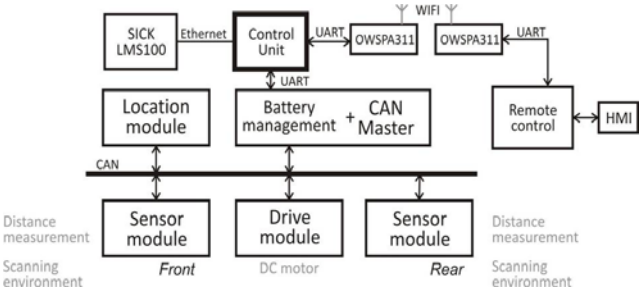


Fig. 9. Architecture of the probe: CAN based distributed control system

During the ride it is necessary to check the state of the embedded power source – the battery. Power consumption of the probe is displayed in (Fig. 7). During the ride the average power consumption of the probe is approximately 2.3 Ampers. The capacity of the battery is 4.6 Ah. So the probe can work in an active state for a maximum of 2 hours. The solar panel is prepared for charging the battery during the ride for the future. In the case of low energy the probe will stop, turn off the sensor system and wait to charge the battery [7].

4.2 SW Architecture of Main Control Unit of the Probe

The main Control unit of the Probe uses a similar SW platform to the remote control. The main control application uses several cooperation processes. This method enables the main control system to dynamically start, stop or replace part of the application without influencing the rest of the application.

5 Conclusion

The proposed set of mobile devices in the article successfully works at the Technical University of Ostrava. The main purpose of the project is to give the tool to the user which can extend the user perspective. The mobile device is equipped with an algorithm for finding the best way from the current to desired position. The outdoor precision obtain from GPS is sufficient, but the precision of indoor localization using only WiFi is insufficient. The current state of the project due to low precision is to use information from distance and other sensor to precise the localization. As a result the current work is focused on using data from laser scanner. In the future dynamic generation of the map based on information from sensors will be implemented.

Acknowledgements

The work and the contribution were supported by the project FR-TI2/273 “Research and development of progressive methods of long-distance monitoring of physico-mechanic quantity including wire-less transmission of processed data.” and the SGS project SP/2011. This work was also supported by the Ministry of Education of the Czech Republic under Project 1M0567.

References

1. Mohapatra, D., Suma, S.B.: Survey of location based wireless services. In: IEEE International Conference Personal Wireless Communications, ICPWC 2005, pp. 358–362 (2005)
2. Hui, L., Darabi, H., Banarjee, P., Jing, L.: Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(6), 1067–1080 (2007)

3. Brida, P., Machaj, J., Benikovsky, J., Duha, J.: An Experimental Evaluation of AGA Algorithm for RSS Positioning in GSM Networks. Elektronika Ir Elektrotechnika 8(104), 113–118 (2010)
4. Krejcar, O., Cernohorsky, J.: Database Prebuffering as a Way to Create a Mobile Control and Information System with Better Response Time. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part I. LNCS, vol. 5101, pp. 489–498. Springer, Heidelberg (2008)
5. Kotzian, J., Konecny, J., Prokop, H., Lippa, T., Kuruc, M.: Autonomous explorative mobile robot: Navigation and construction. In: Proceedings - 9th RoEduNet IEEE International Conference, RoEduNet 2010, art. no. 5541599, pp. 49–54 (2010), ISBN: 978-14244-7335-9,
6. Srovnal Jr., V., Machacek, Z., Hercik, R., Slaby, R., Srovnal, V.: Intelligent Car Control and Recognition Embedded System. In: Proceedings of International Multiconference on Computer Science and Information Technology, IMCSIT 2010 – RTS 2010, Wisła, Poland, October 18-20, pp. 831–836 (2010)
7. Kotzian, J., Srovnal, V.: Distributed embedded system for ultralight airplane monitoring. In: ICINCO 2007, sborníku, Anger, France, vol. 1, pp. 448–451 (2007) ISBN 978-972-8865-82-5
8. Evennou, F., Marx, F.: Advanced integration of WiFi and inertial navigation systems for indoor mobile positioning. Eurasip Journal on Applied Signal Processing (2006)
9. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, Article ID 802523, 8 pages (2010)

Proactive User Adaptive Application for Pleasant Wakeup

Ondrej Krejcar and Jakub Jirka

VSB Technical University of Ostrava, Center for Applied Cybernetics,
Department of measurement and control,
17. Listopadu 15, 70833 Ostrava Poruba, Czech Republic
Ondrej.Krejcar@remoteworld.net, Jakub.Jirka@vsb.cz

Abstract. Paper describe a solution to solve a problem of unpleasant morning wakeup of people by developing of a special application for mobile smart phones (or mobile devices) which provide a smart pleasant wakeup based on detection of users sleep stages. Developed application named wakeNsmile present a Proactive User Adaptive System, which react by his functionality to user request in sense of pleasant morning wakeup. This user request is solved by use of sleep stages detection during the night (sleeping process). These detections running during monitoring phase 30 minutes before requested wakeup [1], [2]. If application detects usefull sleep stage during this time window, the user is wakeup before requested time. If however no detection is happened, user is wakeup at requested time. User adaptivity can be enhanced by user data recording and processing to reach higher level of successful sleep stage detection [3]. When used in connection with EEG signal processing and motion detection of user a complex system can be developed as Proactive User Adaptive System for HomeCare. Such complex solution can improve user comfort (e.g. children don't need further to stay in hospital to monitor their EEG to detect anomalies).

Keywords: Proactive; User Adaptive System; FFT Analysis; Mobile Device; Localization.

1 Introduction

Sleep is a complex process regulated with our brain and as such is driven by 24 hour biological rhythm. Our biological clocks are controlled by chemical substances that are mostly known to us. One of them is hormone called melatonin which is suspected to make us feel sleepy.

This substance is produced in our brain and some scientists believe that it is also cause of a metabolism slowdown before falling asleep. Melatonin secretion leads to the body temperature reduction, blood flow towards brain limitation and muscles slackness.

Approximately two hours after we fall asleep our eyes start to move back and forth irregularly. Based on this fact, sleep stages are divided into two main stages REM sleep with (Rapid Eye Movement) and NREM sleep stage (Non Rapid Eye

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 472–481, 2011.

Movement). NREM sleep is divided into another four sub-stages, when with increasing number the sleep is more and more deeper. [4]

During healthy individual sleep, REM and NREM stages changes a few times. Most of the dreams are happening in REM stage. Body muscles are completely loosened and thanks' to this fact one is awaken refreshed.

During deep (NREM 3 and 4) sleep stages blood pressure is decreasing which lowers chance of cardiovascular danger [4] and growth hormone is produced in its maximum in adolescent age. [5]

Sleep stages:

a) Wake (Awake)
b) REM – we dream in this stage
c) NREM1 – falling asleep
d) NREM2 – light sleep
e) NREM3 – deep sleep
f) NREM4 – deepest sleep

A graph that visualizes sleep architecture in a time (Series of individual stages) is known as Hypnogram. Typical hypnogram is displayed on image (Fig. 1) (duration - 8 hours, obvious periodical changes of deep sleep stages and REM sleep stages).

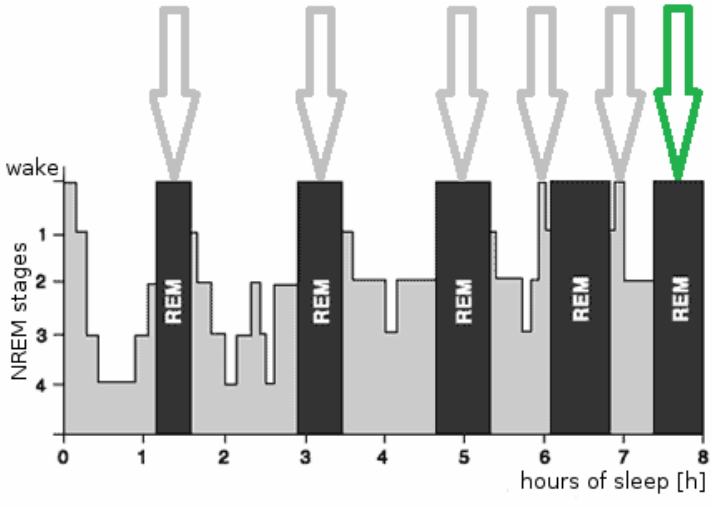


Fig. 1. Hypnogram with optimal wake up periods

We can observe periods of mild sleep as potential candidates to wake up on previous figure. According to information provided in previous paragraphs, the most optimal time period to wake up is of course after 8 hours of sleep. This period is marked by green arrow on picture (Fig. 1).

2 Discussion of Similar Solutions

The final goal of this project is a developing of proactive user adaptive system for pleasant wakeup, which will be operated on windows mobile devices. This UAS application, will implement algorithms to detect sleep stages. Currently there are some similar projects available to download and run onto various mobile device platforms from Nokia, Apple iPhone, Windows Mobile etc. However nobody provide info what algorithms are implemented and how. What level of successfully detected sleep stages and finally waked up users is possible to achieve.

First project from [6] is available to Nokia and iPhone mobile platforms named HappyWakeUp. This solution does not try to wake you up when you are in deep sleep or if you are in REM sleep (sleep phase with dreaming). These are the moments when it is most difficult to wake up [6].

HappyWakeUp application is based on principle of active monitoring of users sleep using the microphone of the mobile phone. It is trying to detect user movements in a bed. This is the reason, why user's mobile device (mobile phone) must be placed near to user. Application then makes statistical analysis of the quality and cycles of user sleep. As authors describe, application is developed to detect only statistically significant movements as arousals, because during these moments users are actually awake or almost awake [6]. Authors however don't publish any other details like algorithms or techniques, so it is impossible to evaluate their software according these principles.

It is also possible to find other similar projects like HappyWakeUp. For example at Macjek Drejak Labs develop a Sleep Cycle mobile application. Their application *Sleep Cycle has become a huge success with a #1 paid app position in many countries, including Germany, Japan and Russia* [7]. Instead of HappyWakeUp a Sleep Cycle application monitor users move by using of embedded accelerometers which are actually equipped with modern mobile devices (like iPhones or some HTC types). User make during different sleep phases different moving in bed, which is possible to determine by Sleep Cycle application. Remainder phases of signal processing stay without any changes [8].

Presented problems with availability of any information about quality of detection or statistical results leads to need of a proactive UAS for pleasant wake up, which is discussed in rest of this paper. Following parts deal with an algorithms and their implementation to real application to detect sleep stages by using of microphone input. Using of accelerometers will be investigated in near future.

3 Implementation of Proactive UAS for Pleasant Wakeup

Our proactive UAS for pleasant wakeup was named as wakeNsmile application (Figure 2). This application is written in C# programming language using .NET Compact Framework in version 3.5, which is a special framework solution from Microsoft for mobile devices [9]. Application was developed in Visual Studio and tested on a HTC Roadster mobile device with Windows Mobile 6.5 operating system.

wakeNsmile application uses user control called Alarm, that has been created as a part of this project. Application is using Math.NET [10] neodym library for FIR

(Finite Impulse Response) filter design and WaveIn and WaveOut libraries [11] for mobile device sound interface communication.

User control Alarm is reusable component and can be added to a newly created project and adjusted by needs of a programmer. User control is using number of other classes that take care of signal recording and analysis (Fig. 3):

- Recorder class
- Analyzer class
- Wave, WaveIn, WnsWaveIn, WaveOut class
- Player class

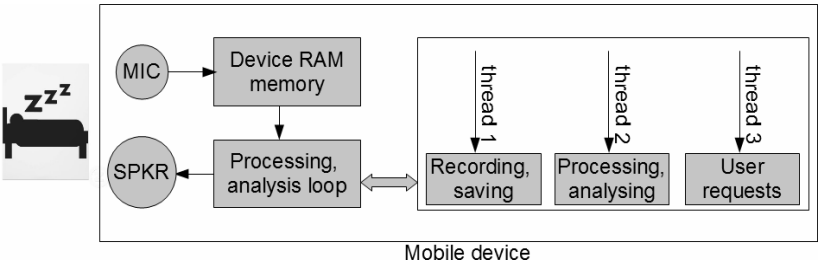


Fig. 2. Block diagram of wakeNsmile application

Recorder, Analyzer and Alarm (user control) classes are all running in their own threads (Fig. 2) so that signal recording is not influenced by signal analysis. By this reason a user is always able to interact with program even during signal recording and analysis.

Recorder class is using WnsWaveIn class [11] and saves the sound recorded from the device microphone input. Class is designed using singleton design pattern because as device can record from only one device input in a time.

Instance of a recorder class is running in its own thread. Thus the graphical user interface Alarm (user control) is ready and available to receive command at any time even during recording and analysis process and Analyzer class is able to analyze data sample at another sample recording without recording interruption.

.NET compact framework does not offer any classes or interfaces to communicate with audio interface on mobile device. Thus WaveIn and WaveOut classes are used. Recorder class contains important constant MAX_REC_TIME, which defines time range of record before it is processed and analyzed. By default it is set to 50ms. That means that 50ms worth data samples are recorded and then handed over to Analyzer class that starts signal analysis in its own thread immediately. This constant can be changed by developer if needed (see end of section 4).

Analyzer class is class intended for recorded signal with MAX_REC_TIME analysis.

Last but not least Player class as wrapper around WaveOut class that plays alarm sound in a loop until stopped by user. Instance of this singleton pattern designed class is running in its own thread as well.

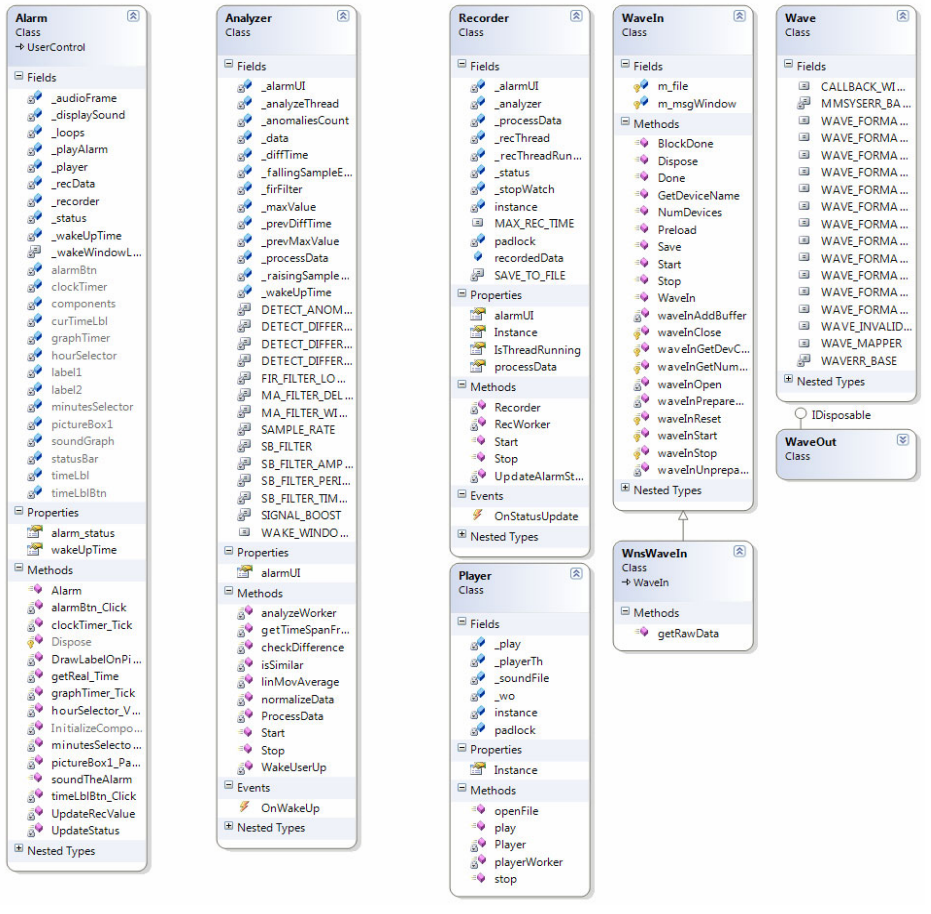


Fig. 3. Class diagram of wakeNsmile application

4 Sleep Stage Analysis Algorithm

Sleep stage analysis algorithm is based upon knowledge about specific sleep stages, and empirical testing on two subjects. Detection algorithm is based on presumption that if patient is in the REM phase and his sleep is not disturbed or pathological in any way his muscles start to spasm that lead into a user movement that itself can be recorded and detected by the device microphone. If patient is in NREM stage 2 and higher sleep stage all of his muscles are relaxed and thus no body movement occurs. [12], [13]. Recording itself is triggered 30minutes (by default) before the wake up time set by the user before sleep. If an erratic movement is detected within this interval user is awaken at that time when the erratic movement occur. Time interval of 30 minutes was estimated from facts mentioned in studies [12], [13]. This time window has represent REM stage of sleep, when we are almost awake. If no movement sound is detected during this 30 minutes window user is awaken on set up time. According to [13], at least once the REM phase occurs during this 30min. window.

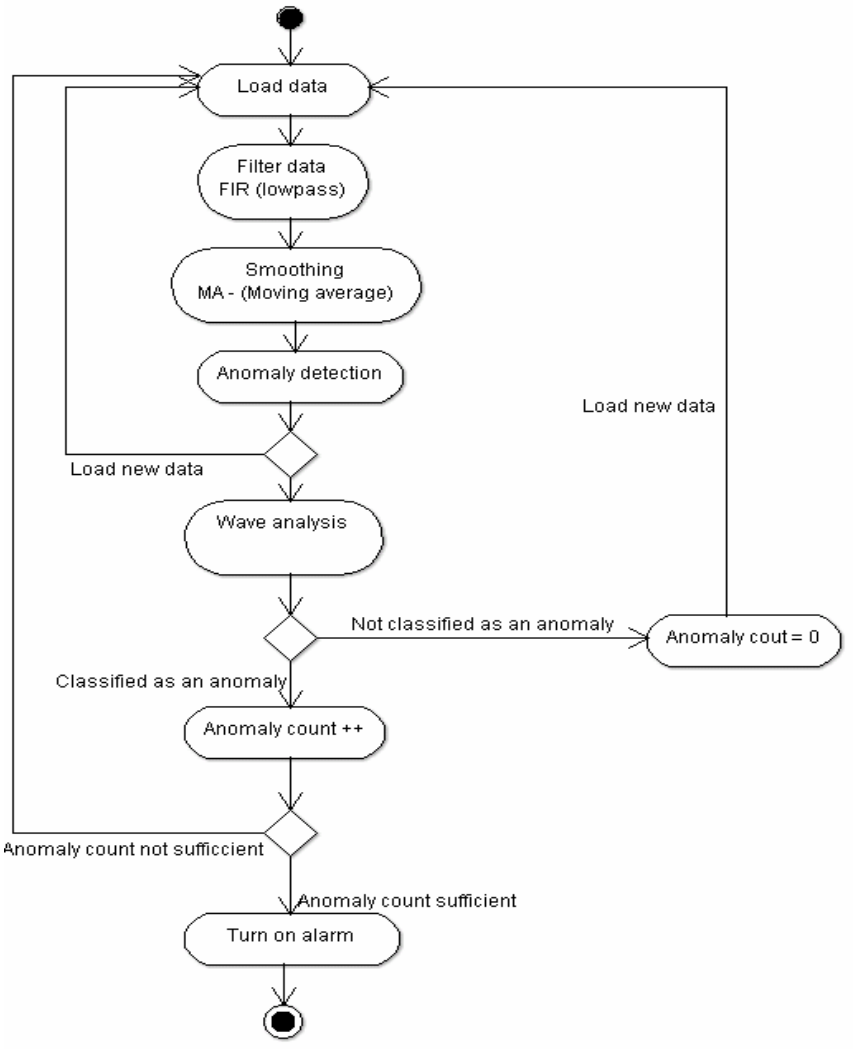


Fig. 4. UML activity diagram sleep stages detection algorithm

Algorithm (Fig. 4) itself is based on time-series analysis. It is based on sound wave differences (period, amplitude), that are abnormal and above isoline. Isoline is computed by linear moving average algorithm. Waves are likely to be more similar within biological manifestations of sleep (breath, snoring). Waves are much more similar in means of period and amplitude. These waves as biological manifestation during sleep are taken as false deviations from isoline as they most likely represent deeper NREM stage of sleep and thus are not positive to wake user up and are taken in final statistical analysis as negative sleep anomaly.

On the other hand if erratic waves (different in period and amplitude) are registered at least DETECT_ANOMALIES_COUNT they are taken into final statistical analysis as positive sleep anomaly. These erratic waves in final analysis are most likely to represent user movement in a bed which is manifestation of brain-motor interconnection that occurs during REM stage of sleep that is very positive to wake user up.

Algorithm described in previous paragraphs is adjustable for developer in many ways. Every constant that can adjust sensitivity of algorithm in any aspect described is shown below. These constants are saved in Analyzer.cs file:

```
//How many minutes before wake up time should detection begin
public const int WAKE_WINDOW_LENGTH = 30;
//How fast signal should change to monitor its tendency
const double DETECT_DIFFERENCE_SPEED = 20;
//Difference between isoline amp. given by MA usually 0
const int DETECT_DIFFERENCE_AMPLITUDE = 40;
//How many s mic input should be above
const double DETECT_DIFFERENCE_SECONDS = 0.1;
//How many anomalies with respect to detect setting there should
const int DETECT_ANOMALIES_COUNT = 3;
//Filter out snore and breathing as periodic biological event
const bool SB_FILTER = true;
//How many periods of SB peaks to wait for
const int SB_FILTER_PERIOD_NUM = 3;
const double SB_FILTER_AMPLITUDE_SIMILARITY_COEFF =
                                DETECT_DIFFERENCE_AMPLITUDE / 4;
const double SB_FILTER_TIMESPAN_SIMILARITY_COEFF =
                               DETECT_DIFFERENCE_SECONDS / 4;
//How big moving average window is (in n. of samples)
const int MA_FILTER_WINDOW_SIZE = 100;
//How big moving average window is (in n. of samples)
const int MA_FILTER_DELTA = 20;
const int SIGNAL_BOOST = 4;  //Amplification of incoming signal
const int SAMPLE_RATE = 22050; //Do not change this - yet
const int FIR_FILTER_LOWPASS_CUTOFF = 2000; // In Hz
```

5 Proactive User Adaptivity

wakeNsmile application is developed to react on users declared request in form of happy wake up at predefined time (Fig. 5). The time defined for alarm is however the latest possible time to wake up of user. We are trying to detect a body state in which the user is most able to wake up with a smile. Time period for detection analysis of state phases is declared to 30 minutes. A Fast Fourier Transformation (FFT) and some other sophisticated methods are used for it. Created application is an ideal example of user adaptive solution for mobile devices. Currently a single application is developed, but a distributed architecture version with a neural network analysis and people

database is planned for future steps to be a completely embedded solution at Mobile UAS Framework. Developed application act as a proactive solution is sense of wake up of users in most suitable time.

Fig. 5. wakeNsmile application example

6 User Adaptive System for Homecare

Developer wakeNsmile application is practicable as it is as an example of proactive UAS for pleasant wakeup. Due to the used modular conception of developer solution, a system is easy extensible and reconfigurable in other UAS solution. The most valuable part of describet solution is Alarm component, which make it possible to use in more complex solution of User Adaptive System for HomeCare. Within this UAS Framework the most usable application area is in home care of young patients (e.g. for monitoring EEG for epilepsy detection). This use is one of further steps in our future research of developed solution.

The goal of this UAS HomeCare Framework is developing of monitoring as well as supervising part of interconnect system for mobile devices (typically mobile smart phones). Monitoring device will use the embedded accelerometer device for motion detection of user. If some motions were detected (e.g. in case of patient/user shaking), a recording of sound as well as EEG monitoring is activated along with particular analysis of these data. In same time alarm is violated at supervisor mobile device (in home care typically parent act for this role) to invoke needed action (e.g. help to patient).

The whole measurement chain along with complete architecture is provided at figure (Fig. 6). From remote PDA of supervisor a next alarm is violated into Central monitoring statio, where paramedics can examine record and make further actions (e.g. Ambulance service invocation to safe patient life in the worst case).

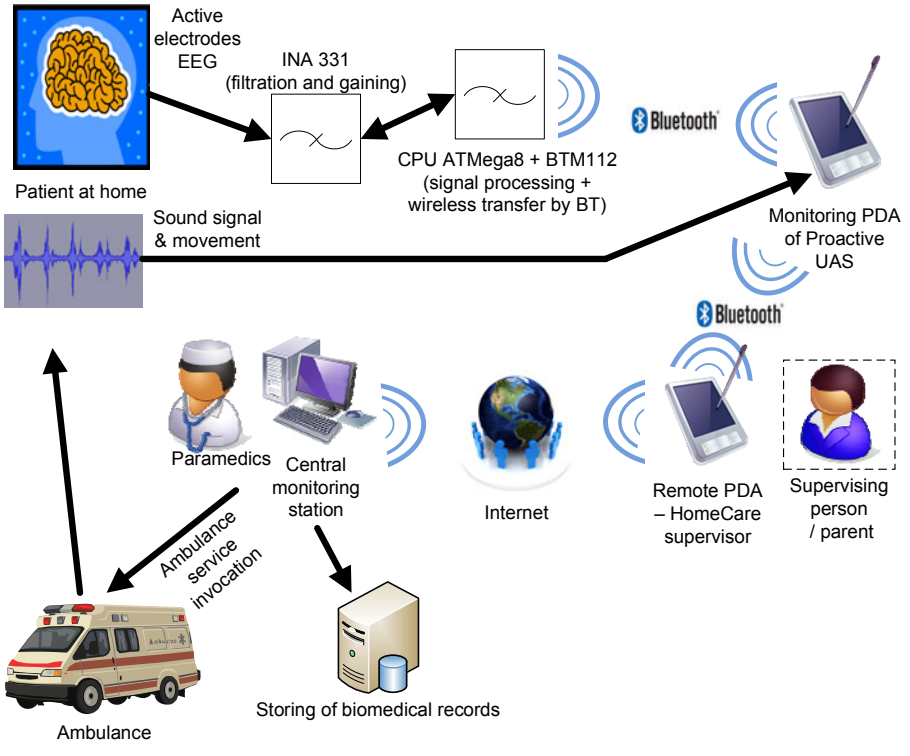


Fig. 6. Measurement chain of EEG, sound and motion activity monitoring. Complete Proactive User Adaptive System for HomeCare.

7 Conclusion

wakeNsmile application was created as Proactive User Adaptive System solution for pleasant morning wakeup of users. Our tests provide more than 72 % of successful happy wakeup from all test persons (over 89% in case of healthy persons).

The solution is working fine, but in some areas where detection algorithm does not work the use is not recommended. The used method of sleep stage detection (sound analysis by use of mobile device microphone) is not applicable for people with sleeping disorders like insomnia, etc whose sleeping manifestations are pathological. This method is also not applicable in noisy areas which are also unhealthy for sleeping. Another problem is physiological manifestations of sleep like snoring or cough. But these manifestations can be detected and filtered out. Snoring itself is physiological periodic process that can be detected and distinguished from erratic movement anomaly sounds that are key data for the detection algorithm. Snoring is an unwanted event that occurs during NREM sleep stages [14], [15].

Acknowledgment. This work was supported by the Ministry of Education of the Czech Republic under Project 1M0567.

References

1. Krejcar, O., Jirka, J., Janckulik, D.: Proactive User Adaptive System for Windows Mobile Devices – Processing of Sound Input Signal for Sleep State Detection. In: ICMEE 2010, Kyoto, Japan, August 1-3, pp. 374–378 (2010)
2. Krejcar, O., Jirka, J.: Design, Implementation and Testing of Mobile Phone Application for Pleasant Wake Up. In: DELTA 2011, Queenstown, New Zealand, January 17-19, pp. 242–247 (2011)
3. Gentle Alarm iPhone Application developed by Craft mobile company (2010), http://gentle-alarm.com/
4. Carskadon, M.A., Acebo, C., Richardson, G.S., Tate, B.A., Seifer, R.: An approach to studying circadian rhythms of adolescent humans. Journal of Biological Rhythms 12, 278–289 (1997)
5. Carskadon M.A.: Delayed Sleep Phase Syndrome in Adolescents, doi: 10.1007/978-1-59745-115-4_6
6. HappyWkaUp smart alarm clock application for Nokia mobile phones (2010), http://www.happywakeup.com/en/
7. Sleep Cycle mobile application, http://www.mdlabs.se/sleepcycle/
8. Phan, J.: WakeMake Analyzes Your Sleep Cycle to Wake You Up Refreshed, http://www.sync-blog.com/sync/2010/01/analyze-your-sleep-cycle-wake-up-refreshed.html
9. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, Article ID 802523 (2009), doi:10.1155/2009/802523
10. Math.NET documentation, OpenSource .NET Projects 2002 - 20010, http://mathnet.opensourcedotnet.info/doc/
11. Recording and Playing Sound with the Waveform Audio Interface, Seth Demsey – Microsoft (January 2004), http://msdn.microsoft.com/en-us/library/aa446573.aspx
12. Fuller, P.N., Gooley, J.J., Saper, C.B.: Neurobiology of the Sleep-Wake Cycle: Sleep Architecture, Circadian Regulation, and Regulatory Feedback. Journal Of Biol. Rhythms 21, 482–493 (2006)
13. Rechtschaffen, A.: Current perspectives on the function of sleep. Perspectives in Biological Medicine 41, 359–390 (1998)
14. Snoring during NREM sleep: respiratory timing, esophageal pressure and EEG arousal
15. Scholle, S., Schäfer, T.: Atlas of states of sleep and wakefulness in infants and children (Somnologie - Schlafforschung und Schlafmedizin), vol. 3(4), pp. 163–241 (1999)
16. Brida, P., Machaj, J., Duha, J.: A Novel Optimizing Algorithm for DV based Positioning Methods in ad hoc Networks. Elektronika Ir Elektrotechnika (Electronics and Electrical Engineering), 1(97), 33–38 (2010) ISSN 1392-1215

Analysis and Elimination of Dangerous Wave Propagation as Intelligent Adaptive Technique

Zdenek Machacek

Department of Measurement and Control, FEECS, VSB – Technical University of Ostrava, Ostrava, Czech Republic
zdenek.machacek@vsb.cz

Abstract. This paper focuses on the intelligent adaptive user system for analysis and detection of dangerous wave propagation with wide setting of parameters and limits, which depends on user type and requirements. The designed mathematical model as user adaptive interface is powerful instrument for complex analysis and approximate result of unhealthy effects of acoustic, and vibration wave's propagation without necessity to practice testing of health parameters. The solution is based on species interaction given by medical and physics regularities. The model makes it possible to intelligent customize various additional depending elements and parameters to the model, with their specification, which is given for example by vibration dampening capacity, noise attenuation, and source power. The paper paragraphs describe a mutual parameter dependences, algorithms, methods and analyses of designed adaptive model with the function mathematical block description. The developed model is constructed in the mathematical program Matlab.

Keywords: user modeling, adaptation, wave propagation, acoustic, vibration, analysis.

1 Introduction

There is presented user adaptively setting system for mathematical modeling and simulation of a wave sources, a wave propagation track, and a wave analyses for unhealthy verification. The developed analysis system is adaptive for various parameters and situation cases. The model parts of whole adaptive analysis system are simulated as particular blocks in Matlab Simulink programming environment by means of structural programming. These parts can be connected and mutually combined to the schemes corresponding to practice situations, which are approach to real unhealthy risks. The mathematical models of named model parts are based on physic formulas, which are solved by means of acoustic and vibration principles as stated below. The equations and mathematical relations are solved by regular numerical methods and blocks programmed in Matlab Simulink, which enables these manual or automatic simulation settings: variable dynamic step, relative and absolute tolerance, type of simulation character.

Due to the extreme range and complexity of this problematic, the presented paper only describes the basic necessary model parts of the whole analysis system for

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 482–491, 2011.

unhealthy wave propagation, which are a wave source block, a wave propagation track block, and a wave analysis block. It is also presented for comparison of simulation blocks implementation and mathematical formulas with defined effects. There is described a basic implementation methods and a main structure of the presented wave propagation model. Last but not least, the paper not only shows the model of a separate model parts itself with defined functionality, but also its integration into the model system with complex wave propagation algorithms, together with wave source, and wave analyses results, as it can be seen as structure in Fig. 1. The solution advantage is that there are many configurations and various parameters settings of each block, which is implemented to the model structure. The interconnections of model blocks differ from case to case; it depends on focused wave analysis and wave source. The following paragraphs deal with description of wave analysis system and particular model block analyses.

The mutual dependencies among the function blocks and mathematical formulas are programmed specially for result of health risks and unhealthy effects. The model structure is implemented by sub-models, which define the mathematical function blocks. The described algorithms inside sub-models were programmed as the Simulink blocks and Embedded-functions combination. An advantage of Embedded-functions is the possibility to construct a basic blocks, which could be used many times in a model with varying parameters. The Embedded-functions are programmed for mathematical set of equations and graphical animations. The presented solution is seriously helpful instrument for design and analyses, which defines unhealthy effects of acoustic, vibration wave propagation, and their reduction compare to unhealthy limits, which are given by available sanitary standards. The model realization enables to adaptively build various situations with defined wave sources, distances, and barriers. Afterward, the results from model computing could be gravely useful for design of barriers, materials, distances, which should be implemented in practice for elimination of unhealthy dangerous wave effects to human body.

2 Mathematical Modeling Structure

The mathematical program environment Matlab with Simulink toolbox provides a powerful instrument for the capabilities of the developed model represented wave propagation and its healthy analyses. This paper describes what the model simulates and how and why it is convenient to use one.

The implemented model structure is divided to sub-models, which represent different functions of the mathematical blocks. The basic sub-models of unhealthy wave propagation modeling are compound by connection or their combination of the wave source block, wave propagation track block, and wave analysis block. The model contains Embedded-functions, which are additional element of Matlab modeling. The Embedded-functions enable to add customized algorithms to Simulink models, either written in Matlab programming language. After Embedded-function has been written and placed its name in an model block, there is possible to customize the user interface by using masking [6],[7].

The sub-models and Embedded-functions can be effectively used for a variety of applications, such as general purpose blocks in Simulink, describing a system as a mathematical equations, graphical animation using. An advantage of sub-models using is that it is possible to build a general algorithms in block that is usable for many times in a model with varying parameters and different purpose. These editable parameters and values are defined by user in each sub-system block setting window [3],[4].

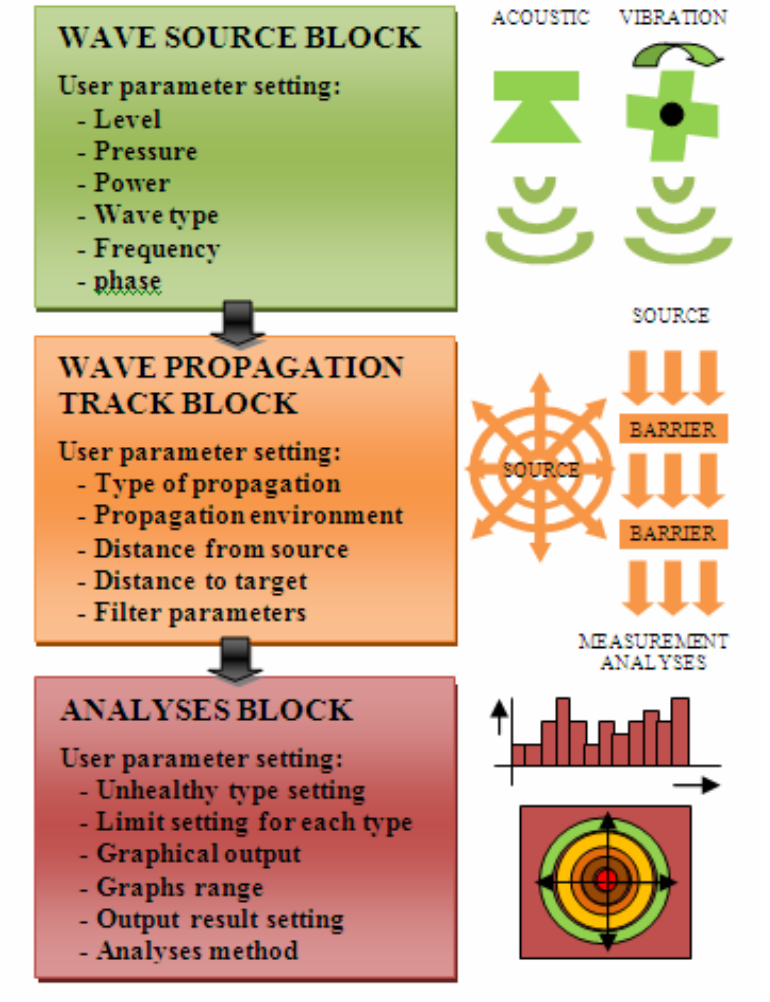


Fig. 1. The basic model structure of adaptive user system for unhealthy wave propagation analyses

3 Simulation of Wave Propagation

The presented simulation comes with purpose of evaluation for unhealthy wave propagation recognition in area with known wave source and barrier. Mathematical algorithms, which are implemented to simulation of wave propagation, are based on combination an equation of motion and on a wave equation. The equations are given by sum of dynamical physical values changes by waves and of steady state of system. An equation of motion for gas and liquid denotes pressure p changes in time t depend on vector of element speed $\overline{v}_i$ changes and capacity power $G_i = 0$, which is equal to zero in this case as shown in the following expression [2]:

$$-\frac{1}{\rho}\cdot\frac{\partial p}{\partial \overline{x}_i}-\frac{\partial \overline{v}_i}{\partial t}+G_i=0, \tag{1}$$

where ρ is liquid density. The vector of position $\overline{x}_i$ could define area from one-dimensional to three-dimensional axes. The wave equation is given by pressure p changes in time t with wave propagation speed c and vector of element speed $\overline{v}_i$ changes depend on position vector $\overline{x}_i$ and liquid density ρ as shown in the following expression [2]:

$$c^{-2}\cdot\frac{\partial p}{\partial t}+\rho\cdot\frac{\partial \overline{v}_i}{\partial \overline{x}_i}=0 \tag{2}$$

The wave source was modeled as a block for acoustic or vibration signal generating with various parameters computation which characterize basic physical values of signal, as signal amplitude W_{MAX}, signal frequency $f[Hz]$, pressure level $L[dB]$, intensity $I[W\cdot m^{-2}]$, pressure $p[Pa]$, power $P[W]$. The generated signal is presented as a harmonic signal with defined outputs which are connectable to the wave propagation track block or the wave analysis block, which is shown as structure in Fig. 2 and Fig. 3. [5].

The wave propagation track is separate to the block for acoustic or vibration signal transmission with various types of medium and character. There is possible to signal propagation as one a dimensional wave, *a* reflection wave, and *s* space wave for various adaptable environment parameters. The graphical results of model characteristics are presented in Fig. 4 and Fig. 5 as wave pressure and spherical propagation depends on distance and time.

Computed values for one dimensional propagation which characterize propagation situation are implemented by the following equations [1],[2]:

- speed potential Φ

$$\Phi = W_{MAX}\cdot\cos(\omega\cdot t-k\cdot x+\varphi) \tag{3}$$

- pressure p

$$p = -\rho \cdot \frac{\partial \Phi}{\partial t} = \rho \cdot \omega \cdot W_{MAX} \cdot \sin(\omega \cdot t - k \cdot x + \varphi) \tag{4}$$

- speed v

$$v = \frac{\partial \Phi}{\partial x} = k \cdot W_{MAX} \cdot \sin(\omega \cdot t - k \cdot x + \varphi) \tag{5}$$

- intensity $I[W \cdot m^{-2}]$

$$I = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} p(t) \cdot v(t) \cdot dt = \frac{p^2}{\rho \cdot c}, \tag{6}$$

where wave number k is given by formula:

$$k = \frac{\omega}{c} = \frac{2 \cdot \pi}{\lambda} \tag{7}$$

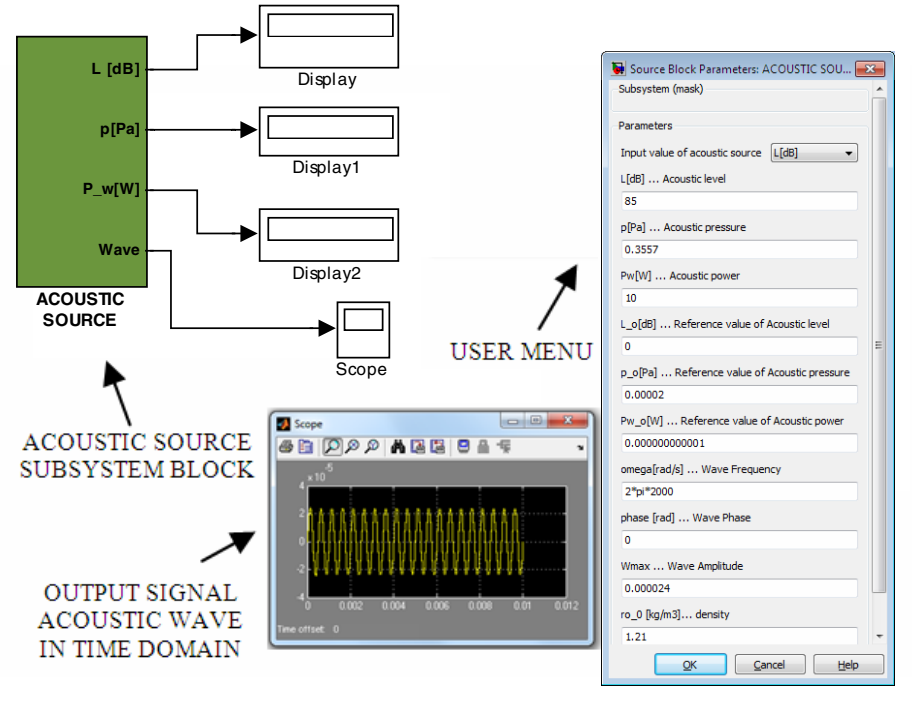


Fig. 2. The block of acoustic source and parameter menu for acoustic source setting

The barriers, which can be inside propagation track, are implemented by sound-proof value $R_\theta[dB]$, for example homogenous soft wall of low frequency signal as shown in the following implemented formula:

$$R_\theta = 10 \cdot \log\left\{1 + \left(\frac{\omega \cdot m}{2 \cdot \rho \cdot c} \cdot \cos\theta\right)^2\right\}, \tag{8}$$

where m is weight of wall and θ is an angle of wave incidence.

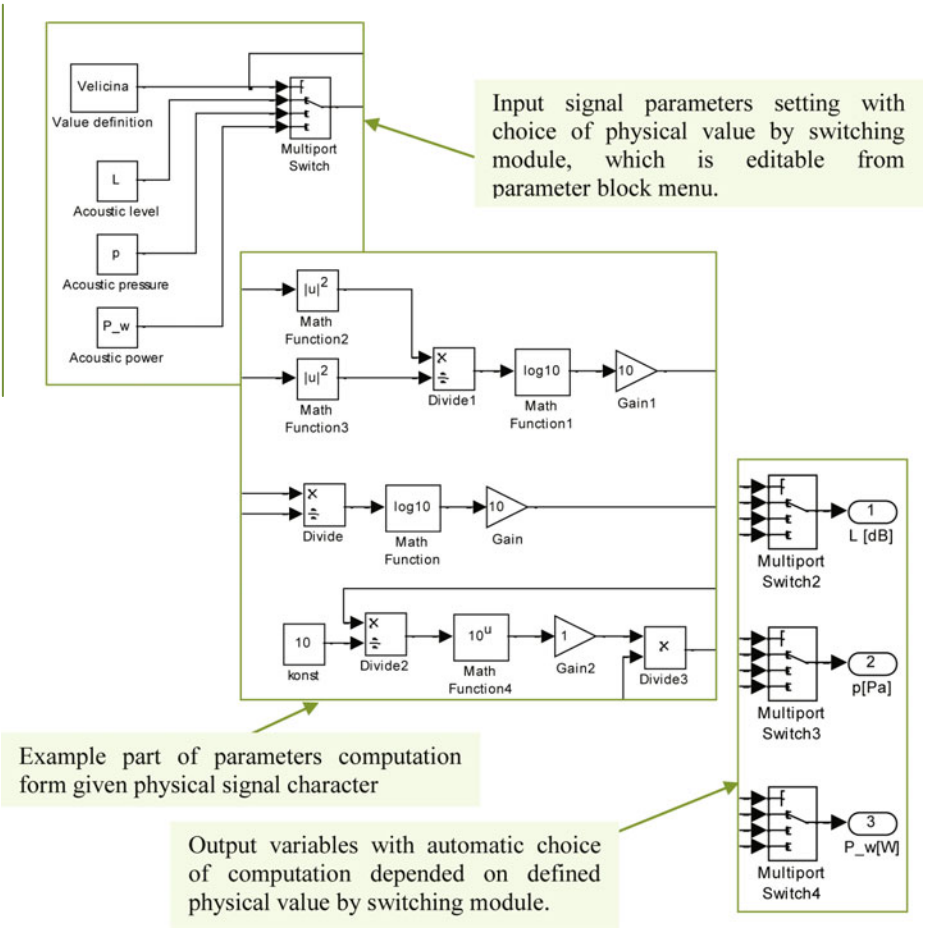


Fig. 3. The example of mathematical model parts of acoustic source, which is developed in Matlab Simulink programming environment

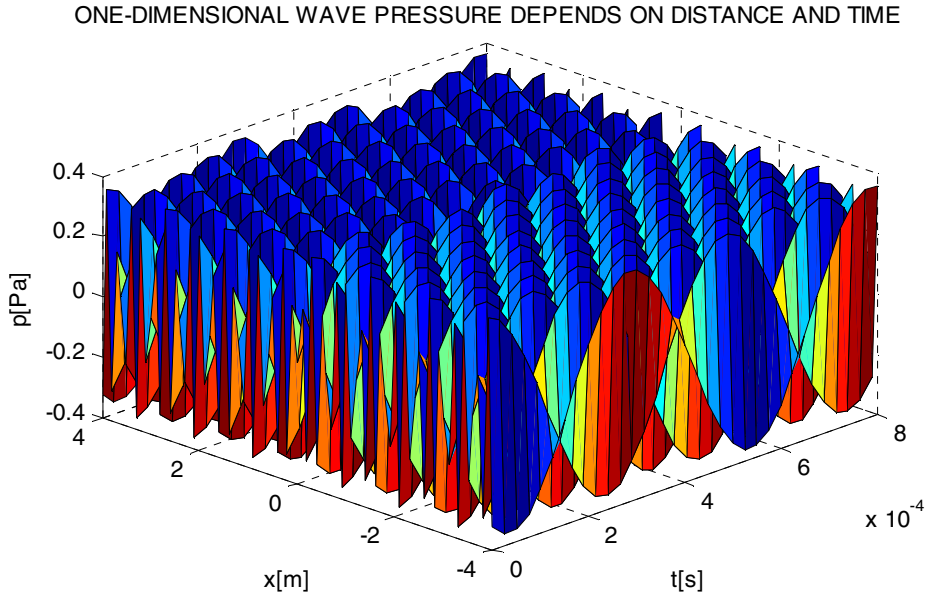


Fig. 4. The graph of wave pressure propagation depends on distance and time

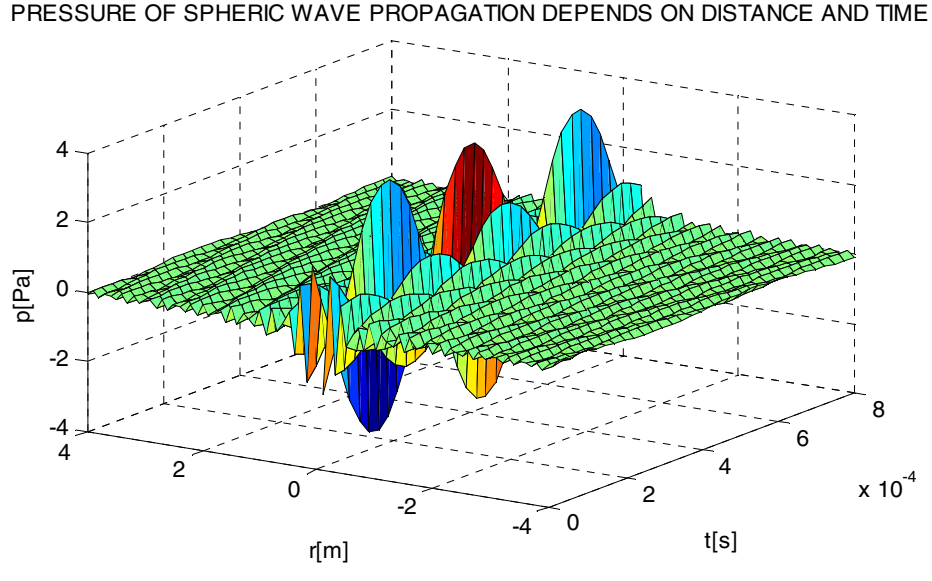


Fig. 5. The graph of wave spherical propagation depends on distance and time

4 Analyses of Unhealthy Waves

The wave analyses are implemented in the block for acoustic or vibration signal recognition and comparison to the various sanitary limits. The results are presented as graphs for signal trajectory showing, frequency spectrum of signal, wave propagation color graphs.

The acoustic propagation analyses, which are simulated, are usable for computation of velocity of wave propagation, sanitary verifying and other parameters by graphical presentation. For sanitary analyses, there is important acoustic wave value of actual acoustic pressure level $L[dB]$, which is presented by implemented equation [1]:

$$L = 10 \cdot \log\left(\frac{p}{p_0}\right)^2 = 20 \cdot \log\left(\frac{p}{p_0}\right), \tag{9}$$

where reference pressure is $p_0 = 2 \cdot 10^{-5}[Pa]$.

This parameter is subjectively feels by human ear in dependency on frequencies from 20Hz to 20kHz, where frequency around 4kHz is the most sensible. The sensitive characteristics-phones are implemented as a collection of interpolated polynomial equations, given by example of equation for the lowest hearing level for frequencies from 1kHz to 16kHz and presented in Fig. 6

$$\begin{aligned} y = &-7.9 \cdot 10^{-26} x^7 + 3.9 \cdot 10^{-21} x^6 - 7.5 \cdot 10^{-17} x^5 + 6.9 \cdot 10^{-13} x^4 + \\ &- 3.2 \cdot 10^{-9} x^3 + 8.0 \cdot 10^{-6} x^2 + -0.01x + 10.3 \end{aligned} \tag{10}$$

The designed analyses are component of graphical results and they evaluate biological effects on acoustic noise, which are defined by standard specification as time limiting noise, impulse noise, infra-noise, ultra-noise, communal noise and others. The dangerous and restricted levels of noise are marked for recognition.

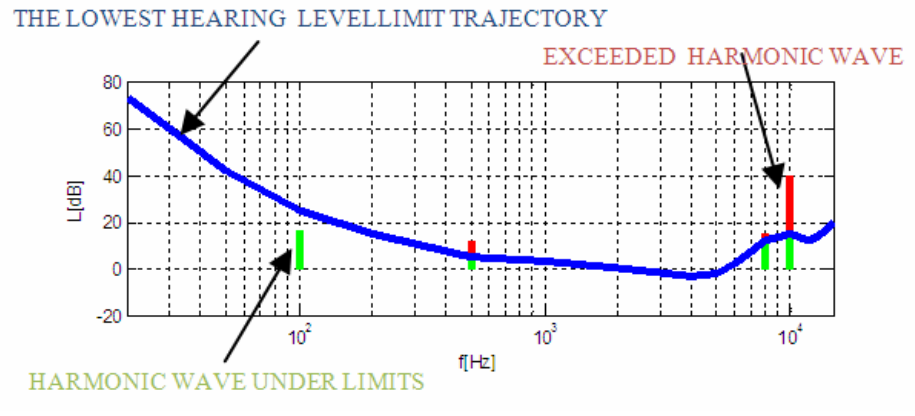


Fig. 6. Waves analysis example in frequency domain

The other wave propagation type in simulation is vibration of mechanical systems. For sanitary analyses, there is important vibration wave value of acceleration level $L_a[dB]$, which is presented by equation [1]:

$$L_a = 20 \cdot \log\left(\frac{a}{a_0}\right), \tag{11}$$

where reference acceleration is $a_0 = 10^{-6}\left[m \cdot s^{-2}\right]$.

The vibration analyses as previous are component of graphical results and they evaluate biological effects on vibration, which are horizontal or vertical, time limiting, defined on part of body and others. The dangerous levels are marked for recognition too.

5 User Adaptation in Mathematical Modeling and Analysis

The applications designed for simulation could support the adaptive capabilities and user expertise mathematical modeling, processed through the user model parameters dealing with the various system responses. The purpose of the described research is almost to investigate how the users could interact a developed program in a more natural way with the aspects influence the program's interaction capabilities and the program naturalness of the dialogue. The developed user adaptive programs have to keep the recommendations of the appropriate reaction and possibility of modeling situation, which has to correspond to physical possibilities.

The solution of adaptable model should be seriously helpful instrument for design and analyses, which defines and simulate characteristic dependences, which are given by available standards and mathematical equations. The model realization enables to build various situations with defined parameters. Afterward, the results from model computing could be gravely useful for design of real application, which could be implemented in practice usage. The mathematical modeling of complex adaptive system should be constructing by basic elementary blocks which represent the mathematical functions, which could be used many times in a model with varying parameters.

6 Conclusion

The purpose of the paper was to verify the possibility of user adaptability in the simulation and modeling for analyses of unhealthy acoustic and vibration waves, which combines theoretical physic knowledge and medical skills. The adaptable model is described by structure in Fig. 1. The particular parts of blocks were developed to theoretical knowledge.

The basic innovation is that the developed wave propagation and analyses user adaptable model. The designed model of analysis wave propagation system is very useful instrument for various adaptable analyses of unhealthy effects to human body of acoustic, vibration wave's propagation. Thanks to the model, there is possible to

reduce effectiveness by defined barriers compare to unhealthy limits, which are given by sanitary standards. The model realization is variable for situations and user given by parameters of wave modules. The adaptively mathematical model is usable for design of barriers, specifying of distances, verifying of measured data, unhealthy factors analyzing. The project simulation results are prepared for verifying by measurement.

Acknowledgement

The work and the contribution were supported by the project Technology Agency of Czech Republic – TA01020282 Enhancement of quality of environment with respect to occurrence of endogenous fires in mine dumps and industrial waste dumps, including its modeling and spread prediction.

References

1. Skvor Z.: Akustika a elektro-akustika (2001)
2. Ziaran S.: Ochrana cloveka pred kmitanim a hlukom (2001)
3. Nevriva, P., Machacek, Z., Krnavek, J.: Simulation of Thermal Fields of Sensors Supported by an Image Processing Technology. In: WSEAS Automatic Control, Modelling and Simulation, p. 7 (2008)
4. Ozana, S., Machacek, Z.: Implementation of the Mathematical Model of a Generating Block in Matlab & Simulink Using S-functions. In: The Second International Conference on Computer and Electrical Engineering ICCEE, Session 8, pp. 431–435 (2009)
5. Vasickova Z., Penhaker M., Augustynek M.: Using frequency analysis of vibration for detection of epileptic seizure (2009)
6. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, 8 (2009), doi:10.1109/EURCON.2009.5167783
7. Tutsch, M., Machacek, Z., Krejcar, O., Konarik, P.: Development Methods for Low Cost Industrial Control by WinPAC Controller and Measurement Cards in Matlab Simulink. In: Proceedings of Second International Conference on Computer Engineering and Applications, pp. 444–448 (2010), doi:10.1109/ICCEA.2010.235.
8. Bencur, A., Smid, J., Kotzian, J., Pokorny, M.: Measurement and modeling in sensor network. In: Proceedings - 9th RoEduNet IEEE International Conference, RoEduNet 2010, Sibiu, Romania, pp. 424–429 (2010), ISBN 978-1-4244-7335-9

User Adaptive System for Data Management in Home Care Maintenance Systems

Marek Penhaker, Vladimir Kasik, Martin Stankus, and Jan Kijonka

VSB - Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Department of Measurement and Control, Ostrava, Czech Republic
marek.penhaker@vsb.cz, vladimir kasik@vsb.cz,
martin.stankus@vsb.cz, jan.kijonka@vsb.cz

Abstract. User adaptive systems for data management becoming more important in the home security and maintenance systems. There does exist plenty of monitoring systems with relevant outputs without right driven information between user and maintenance system. Especially the real time biotelemetry system base on vital function monitoring and first-aid treatment are the main focuses for user adaptation. The goal of this work is design and implementation of such kind of maintenance system providing information by user adaptive system for protection and saving haleness and human life. System design consists from hardware implementation of the sensing and transmission part and from logical structure by software implementation with user adaptation.

Keywords: User Adaptive, Embedded System, Biotelemetry, Electrodes.

1 Introduction

Embedded systems are the combination of hardware and software which are parts of larger arrangements devices. Most of these are part of the control device, so is also known as embedded control system. Embedded systems are mostly designed for user adaptability in operational to respond occurrence in real-time. Those embedded systems may rapidly change the way of human's interaction with their surroundings in conjunction with a number of units and sensors, which can provide the information captured, shared and processed in entirely new ways. Many of the embedded systems were designed mainly according to actual requirement.

The large group of embedded systems is used in medicine, medical devices, self monitoring devices and more and more in health home care applications. With regard to the health-care embedded systems and monitoring devices, valuation and the necessary for more health maintenance monitoring system implementation growing with the number of group of an elderly people and senescent singles. Some of them are familiar wit actual information technologies and they will involve user adaptability embedded systems for vital function monitoring as accessory of daily life as to which can safe their life in critical situation.

The health care embedded systems can be spread onto three extensive categories. These categories are the managing unwell patients with chronic conditions or implants, managing the wholesome people as prevention and finally can be used as clinical support. [1]

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 492–501, 2011.

In health home care systems the patients should check personal health status by themselves with their health status real time feedback displayed. In line with representative biomedical signals that need to be measured from the patients, various types of medical transducers will be connected to the embedded system for signal acquirement. By sensing the discrete biomedical signals to embedded parameter detector, extractor the relevant parameters will be data mined by using promising mathematical techniques. [2]

2 Home Care User Adaptive System Concept

The concept of biotelemetric embedded health supervisory systems is based on usability inside and outside the user flat. There is the mobile part of measuring system which can be taken also outside the flat. The sensor and actuator placed in the flat as fixed or movable or mobile like ECG electrodes connected to the mobile part. All the data from sensors and mobile unit are collected in the home acquisition unit placed in each flat unit. This unit communicates with the surrounding sensors and actuators by wire or wireless way. Out coming communication between flat acquisition unit and dispatching centre is made by Ethernet as Fig.1. [4]

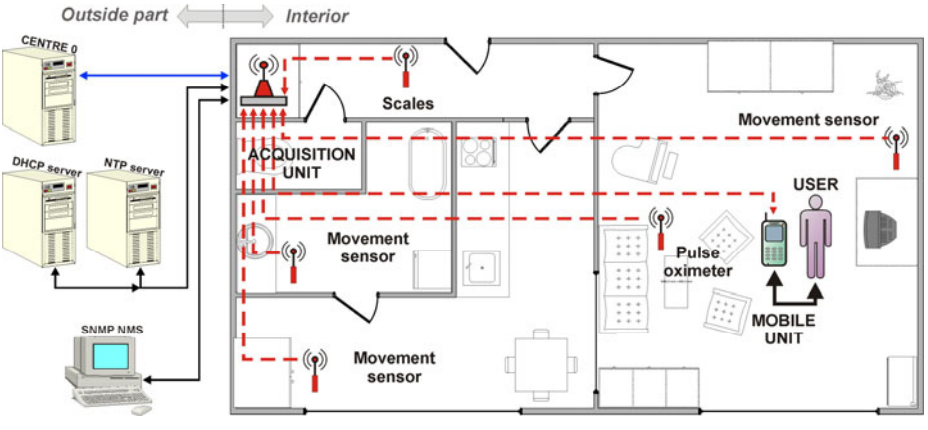


Fig. 1. Home Care User Adaptive Health Supervisory System Concept with Communication Schematic Visualization in Interior and Outer Part

Mobile unit with other sensors can be wearable on the user to take out away from flat. Then after the unit becomes standalone concerning data measuring and transmission by GPRS data connection. In the same time the GPS localization is add into the measured data.

Main property of mobile unit is mobility with light weight, long life battery profile, first aid emergency pushbutton and user interface for self check information. Appearance of mobile unit can be similar to mobile phone due easy user utilization. The sensors and home health embedded systems also have to be untroubled for daily usage with the low cost of arrangement and operation. This system provide reliable, secure and diagnosis prediction credulity respecting of the hardware and software parts. [5], [6]

3 Embedded Hardware

The main hardware components of inner part in Home Care measurement system are mobile part, acquisition unit and set of sensors. The sensors are particular components which are movement and position sensors as static part and personal medical diagnostics systems as semi static sensors. The semi static medical diagnostics sensors like body weight, non-invasive blood pressure, pulse oxymetry, glycaemia, temperature and indoor humidity and lighting. Those sensors provide user interacted data for improvement the diagnosis. Those data can be used for health condition status prediction concerning circadian rhythm and historical data trends. The special part of sensors is mobile wearable sensors placed on body of use and connected by wire into mobile part. The sensors are devices embedding data into biotelemetric embedded health supervisory systems without any requirement for critical data processing performance.

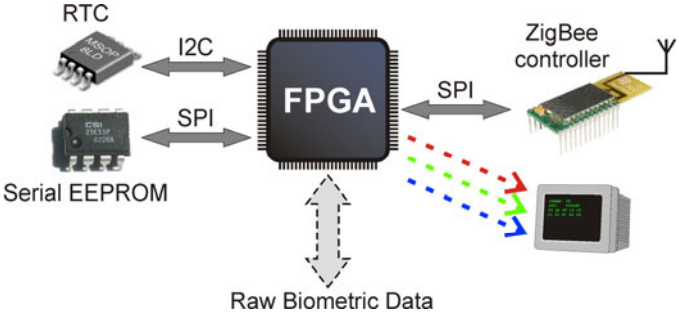


Fig. 2. HW Structure of the Mobile Unit with Programmable FPGA Device. Serial EEPROM is Used for Actual Communication Parameters, ID Codes and the Time Stamps Derived from RTC. RGB Video Interface is Ready for Diagnostics of the Logic Design.

Important role has various management servers and data aggregation points in outside part of system. These points are made of standard server hardware parts and their performance for data effect manipulation is further than appropriate. Key unit for embedded data handling and processing are mobile part and acquisition unit. There is a unification microcontroller design on both units due to the efficiency and cost point of view. Block scheme of this architecture can be seen in Fig. 2.

FPGA device in Fig. 2 constitutes the heart of the mobile unit. That device is used for wide range of concurrently processed operations. The most important ones are listed below:

The first stage of the FPGA design structure is a Raw Biometric Data Parsing, which classifies several data types acquired from sensors. The most important functions utilize DSP features of the FPGA like DSP preprocessing and Harmful Event Detection. All the processed data are evaluated depending on network structure parameters. Next some data frames are marked with time stamp and then the data are encapsulated into frames and compressed. Last logical block adjacent to serial communication pins is an SPI control unit for ZigBee interface. All sequential controllers

inside the FPGA logic are designed as safety synchronous Finite State Machines (Figure 3) with one-hot encoding.

All signal processing functions inside FPGA use specific HW features of the device including DSP cores and dedicated multipliers. Time stamps feature with events saving into EEPROM is helpful in the case the wireless communication should temporarily lost. For that event also the embedded FIFO buffer is needed.

Communication between FPGA and peripherals is implemented using three wire SPI and I2C bus. Both ZigBee controller and FPGA perform their tasks autonomously and can achieve high communication throughput.

Another FPGA function is a logical diagnostics module implemented in the design. For that feature the VGA interface enables real-time monitoring of essential processing parameters of the mobile unit on the common VGA display.

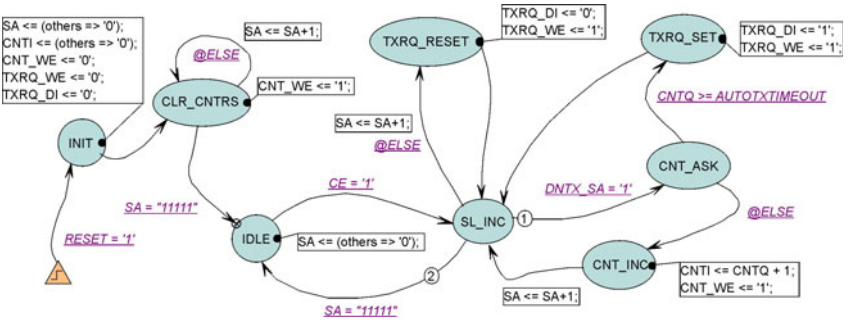


Fig. 3. State Diagram of the Communication Controller for Translation to VHDL Unit. The Finite State Machine (FSM) with 8 States is Controlled by 20 MHz Clock and Utilizes about 23 FPGA Slices.

Differences between unit of acquisition and mobile unit are in their power solution. Mobile unit is mobile device powered by LiIon mAh accumulator with full output operating in three days. Data acquisition unit is mains powered device connected to the Ethernet. For ZigBee wireless communication is used 2.4 GHz ISM band ZigBee chipset working with ZigBee protocol stack. [3]

Caching of biotelemetric data is implemented with 1MB fast FRAM memory. Main goal of cash memory is providing temporal data storage to prevent short time outage on ZigBee or Ethernet network.

Table 1. Device Utilization Summary for Home Care Embedded System Designs

Logic Utilization	Resources		
	Used	Available	Utilization
Slice Flip-Flops	1715	3840	45%
Number of 4 input LUTs	2109	3840	55%
Number of occupied Slices	1195	1920	62%
Number of Block RAMs	10	12	83%
Total equivalent gate count for design		43315	

3.1 Interior Equipment Description

User adaptive system is designed as embedded system implemented in place where the user spends their time like home or rest home. These interiors part primary measure and monitor biological data from patient and handle it through acquisition unit outside the flat thought Ethernet connection. Communication between the interior parts of the monitoring system is realised by short range ZigBee communication technology. There are main components in wireless net like data acquisition unit, mobile unit and sensors as in Fig. 4. In case of larger distance wireless communication between one flat components can be add the ZigBee routers as communication repeater.

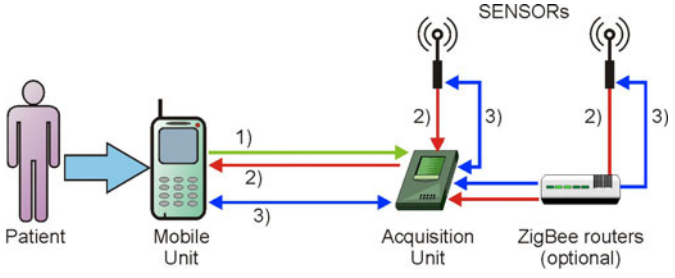


Fig. 4. Communication in Interior Part of Home Maintenance Biotelemetric System. The Data from Sensors Flow through Acquisition Unit to Mobile Unit to User Adaptive System.

Constituent part of this user adaptability system is mobile unit which measure biomedical data from wearable on body sensors like multi-channel ECG, body and surrounding temperature and standing and falling acceleration. The mobile unit is inbuilt into the standard mobile phone case equipped our embedded electronics. Data outgoing from mobile unit and from sensors in interior part are aggregated into acquisition unit in flat and then transferred by Ethernet communication transfer. Data acquisition unit is the only data interconnection part between outside and inside parts. [7], [8]

3.2 Outside Segment of Maintenance System

In case of straight interaction and ambient intelligence in medical systems communication midst healthcare processes and the home care supervisory systems and the profound embedding of wireless sensor into the neighbourhood. These are the most accountable feature of embedded systems application from point of view technological infrastructure for ambient intelligence realization.

Transmitted data from acquisition unit have to be processed, stored and mostly real time analysed for feedback interaction in user or first aid in case of serious disorders. This outside part are mostly dispatching centre that collection multiple interior parts in one time to monitoring individual users. Parts of the maintenance system in interior and outside part handover data through TCP/IP protocols through Ethernet communication transfer as Fig. 5. Outgoing part consist from dispatching centre, servers that providing protocols settings, real time providing servers and Simple Network Management Protocol for managing the communication net.

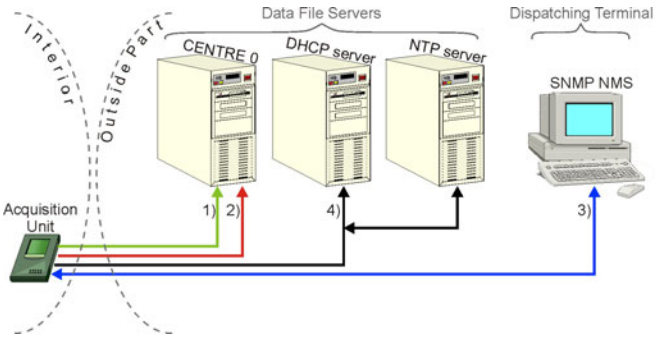


Fig. 5. Outside Part of Home Maintenance Biotelemetric System. The Acquisition Unit Provides Data from Monitoring Unit to Outside Situated Communication with DHCP and NTP Servers. Main Terminal Displays the Actual Status and Alerts from Flat Unit.

4 User Adaptive Interface

Mobile unit is able to monitor basic man's life function. This unit integrates all the measurement, visualisation and wireless communication properties. For the user interface is used LCD display, matrix alphanumerical keyboard and ON/OFF button driven by FPGA with external flash memory for maps and trends history storage. Effect of display and keyboard is admitting easy interaction with user. User can select his measured biomedical data. Keyboard admits customization of graphical user interface on Fig. 7. Sensor and data interface on the mobile unit makes possibility for uploading new user interface.

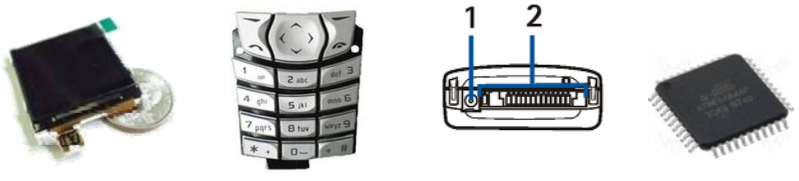


Fig. 6. User Adaptive Interface Components from the Left LCD matrix display, Matrix , Alphanumerical Keyboard, Data Interface with (1) power connector and (2) data connector, FLASH memory for user data storage

Mobile unit is in direct interaction of user. This unit has numerical keyboard, LCD display and several pushbuttons which makes it possible for user to set on values of monitoring, individual pre-sets and visualise measured data Fig. 6.

Data interpretation is realised by graphic user interface on embedded system like mobile phone. The user has straight contact with the mobile unit for visualising actual measured data or pre sets values of monitoring. There are main information on display like battery status, wireless network quality and actual time.

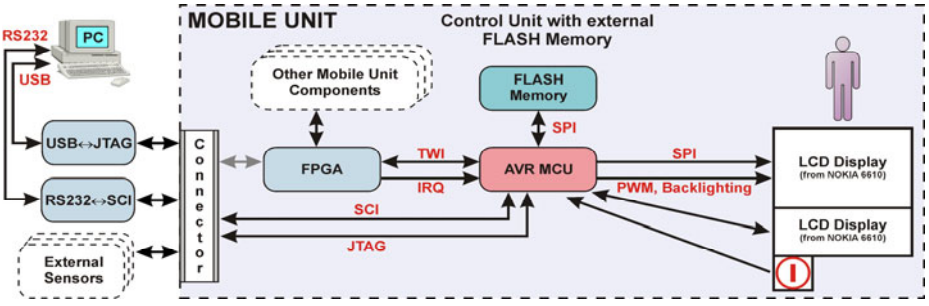


Fig. 7. Main Controlling part of the Mobile Unit is FPGA Device. User Adaptive Interface is driven by microcontroller with AVR architecture. Communication between them by I2C both side communication.

Fig. 8. Initial Screen of User Interface with Main Menu Items

In context switch menu user adjusts the visualisation of alphanumerical data or their graphic visualisation Fig. 8. Possibility of visualisation of electrocardiogram, data trends or blood pressure values is shown in Fig. 9. Inseparable part of user adjusting is settings of signal feed speed, intensity of LCD backlighting etc. There is also push panic button and demo mode for introducing the system.

Actual Measurement Visualization. ECG (channel 1 and 2), Plethysmography (including SpO2 and heart beat). Last twenty minutes of measured data is stored on Flash memory for post processing and pre event analyzing.

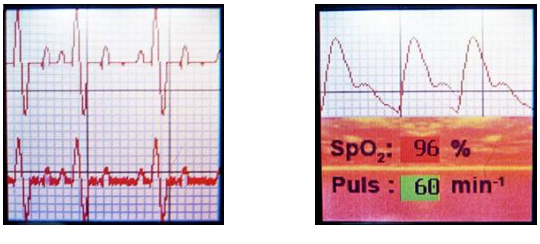


Fig. 9. Records Visualization on User Interface from the Left Real dual channel ECG, Finger Plethysmography

Trends Visualization like body and surrounding temperature, systolic and diastolic pressure, heart beat, SpO2, body weight, position trajectory in room Fig.10. In trends Visualization can be set on the same scale for all kind of data trends.

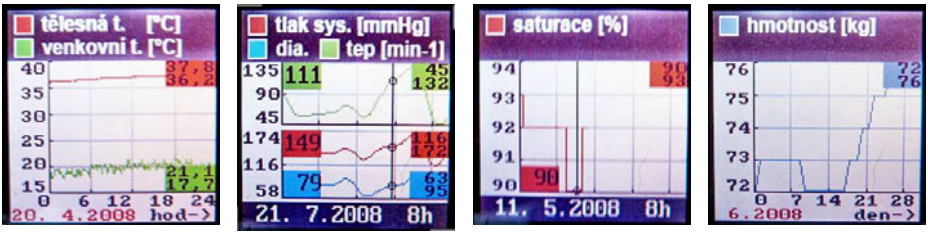


Fig. 10. List of Interacting Trends Chart Accesable from User Interface from thje left Body and Surrounding Temperature, Systolic and Diastolic Pressure, Oxygen saturation - SpO2, Body Weight

Additional Information can user chooses in menu like actual user position in flat can be shown as point on the map Fig. 11. Also the first aid can be used. In menu event visualization can be shown. There is also Flash memory actualization for uploading maps and user interface pre sets values. In Demo mode there is introduction into the system and demonstrate how mange the user interface. Speed of screen feed and backlighting use can adjust by their own needs.

Fig. 11. Additional Informations in User Adaptive System Contains from the Left: User Position, Firs Aid Call, DEMO Mode and Flash Updating Menu

Data that are measured and visualised on mobile unit screen are transmitted by identified ZigBee cluster ID wireless security communication. The same data source and packet transmission identification is used for communication with outer part of maintenance system by Ethernet thought acquisition unit.

5 Data Communication in Health Supervisory Systems

Data communication in biotelemetric embedded health supervisory systems is partially realised in wire and ZigBee wireless data transmission shown on Fig. 4. The main effect of that data transmission is easy monitoring and management on internal

part. Concentration interface between internal and external part represents acquisition unit which is the data interpreter. This unit use SNMP technology for management and internal part of system instatement management by dedicated ZigBee links. There is dedicated 16 bit ZigBee cluster ID for management link between every single ZigBee device and acquisition unit. In data communication are defined messages containing notification of error and their type, commission for system reset, battery charge condition of each elements, timing and ZigBee network settings including PAN identification. There is one to one mapping between data transported by different network technologies without any data and protocol translation. [9]

Data flow controlling is providing in outside part of biotelemetric health supervisory systems without conversion into ZigBee messages of internal part of system. The outside part data communication is driven by DHCP and NTP technology for assigning TCP/IP protocol coherent settings, IP address of SNMP Network Monitoring Station.

6 Conclusion

Nowadays the health maintenance telemetry systems form part of daily life. Everydayness will be in the near future monitored by health embedded systems and inspect our lives and health. In our work was such kind of health maintenance system conception introduced. Presented system is implemented in our two-room testing flat carried on VSB – Technical university of Ostrava and University of Ostrava. Health maintenance system is presently realised for singles indoor monitoring. Nowadays systems implementation provides priceless information about the hardware configuration improvement in the future and also valuable health data files for more accurately diagnosis assesment. In the future we would like to improve the system by integrating GPRS and 3.5G communication for outdoor on-line monitoring including position monitoring by GPS system.

Acknowledgments. The work and the contribution were supported by the project Grant Agency of Czech Republic – GAČR 102/08/1429 “Safety and security of networked embedded system applications”. This work was supported by the Ministry of Education of the Czech Republic under Project 1M0567. This work was partially supported by the faculty internal project,“ Biomedical engineering systems VII”.

References

1. Stankus, M., Penhaker, M., Srovnal, V., Cerny, M., Kasik, V.: Security and Reliability of Data Transmissions in Biotelemetric System. In: XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, MEDICON 2010, IFMBE Proceedings, Part 2, Chalkidiki, Greece, May 27-30, vol. 29, pp. 216–219. Springer, Heidelberg (2010), ISBN 978-3-642-13038-0 (Print) 978-3-642-13039-7 (Online), ISSN: 16800737

2. Srovnal, V., Penhaker, M.: Electronic Embeddes System Applications. In: Proceedings of 2nd International Conference on Mechanical and Electronics Engineering, ICMEE 2010, Kyoto, Japan, August 1-3, vol. 1, pp. 394–398. IEEE Conference Publishing Services, NJ (2010), doi:10.1109/ICMEE.2010.5558521, ISBN 978-1-4244-7480-6
3. Farahani, S.: ZigBee Wireless Networks and Transceivers. Newnes (2008)
4. Černý, M.: Movement Monitoring in the HomeCare System. In: Schleger, D. (ed.) IFMBE Proceddings, vol. (25), Springer, Berlin (2009), ISBN 978-3-642-03897-6; ISSN 1680-0737
5. Havlík, J., Uhlíř, J., Horčík, Z.: Human Body Motions Classifications. In: IFMBE Proceedings EMBEC 2008, CD-ROM. Springer, Berlin (2008), ISBN 978-3-540-89207-6
6. Krejcar, O.: Problem Solving of Low Data Throughput on Mobile Devices by Artefacts Prebuffering. EURASIP Journal on Wireless Communications and Networking, Article ID 802523, 8 p. (2009), doi:10.1155/2009/802523
7. Noury, N., Poujaud, J., Lundy, J.E.: Multidimensional context analysis for recognition of health risk situations. The paradigm of fall detection, IRBM 30(5-6), 268–272 (2009), doi:10.1016/j.irbm.2009.10.008, ISSN: 1959-0318
8. Krejcar, O.: Localization by Wireless Technologies for Managing of Large Scale Data Artifacts on Mobile Devices. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 697–708. Springer, Heidelberg (2009), doi:10.1007/978-3-642-04441-0
9. Scanaill, C.N., Carew, S., Barralon, P., Noury, N., Lyons, D., Lyons, G.M.: A review of approaches to mobility telemonitoring of the elderly in their living environment. Annals Of Biomedical Engineering 34(4), 547–563 (2006), doi:10.1007/s10439-005-9068-2, ISSN: 0090-6964

Effect of Connectivity and Context-Awareness on Users' Adoption of Ubiquitous Decision Support System

Namho Chung[1] and Kun Chang Lee[2,*]

[1] Associate Professor
College of Hotel & Tourism Management
Kyung Hee University
Seoul 130-701, Republic of Korea
nhchung@khu.ac.kr
[2] Professor of MIS at SKK Business School
WCU Professor of Creativity Science at Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
Tel: +82 2 7600505; Fax: +82 2 7600440
kunchanglee@gmail.com

Abstract. Recent surge of ubiquitous computing devices enables users to carry their own ubiquitous decision support systems (UDSS) and make important decisions by using them. However, there few studies investigating why UDSS users intend to use the systems continuously because there are a large number of elements to be taken into consideration. Peculiar factors about the UDSS that this study focuses on include the connectivity and context-awareness function, which were not considered in previous studies of the adoption of UDSS. Accordingly, this study empirically explores this research question by using the scanphone-based ubiquitous delivery system(UDS), which many delivery service providers have been adopting recently as a form of the ubiquitous decision support system (UDSS). The results reveal that the connectivity, context-awareness function, and perceived values play meaningful roles in the UDSS. Based on the results, this study suggests implications and directions for future research for planning and realizing the future UDSS.

Keywords: Ubiquitous decision support system(UDSS), connectivity, context-awareness function, perceived value, Ubiquitous delivery system(UDS), scanphone.

1 Introduction

The corporate environments of information technology have been rapidly changing recently. The individualistic work environments have been shifting toward the network-based cooperation systems according to the advent of the internet. In Korea, particularly, such changes in the work environments are accelerating due to the growth

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 502–511, 2011.

of the mobile market, and the so-called 'ubiquitous environment' is arriving. In the ubiquitous environment, it is expected that higher interaction will be accomplished between companies and their clients as well as providers and their partners, and the decision-making methods will also change according to the changes in the work processes and methods [1]. As a result, the roles of the decision support system that supports decision-making in the ubiquitous environment will diversify and attain greater importance. However, although previous studies reviewed the UDSS, they did not perform enough research on the actual acceptance status and the key functions of the UDSS. Toward this end, this study intends to suggest the scanphone-based ubiquitous delivery system(UDS), which is being adopted by numerous delivery service providers recently, as a form of the UDSS and empirically analyze how the UDSS is being accepted by users. This study is to examine the factors that affect the continuous intention to use the UDS of current UDS users. The following are the study objectives: First, this study aims to evaluate the functions of the UDS as a UDSS, defining the required function of a decision support system in the ubiquitous environment as controllability. Second, this study also intends to examine the effects of the connectivity and context- awareness function, which are the key natures of the UDSS, and the perceived values of the UDSS on the continuous intention to use through trust and decision satisfaction.

2 Literature review

2.1 Ubiquitous

Ubiquitous, as Weiser [2] suggested, means a state where computers and objects are interconnected through a network without any spatial and temporal constraints. He advocated the environment where all objects are equipped with microcomputers and connected to a network by adopting the concept of computing everywhere. Weiser [2] interpreted the ubiquitous network as three concepts: Everywhere-on, whatever-on, and always on. In addition, use of the terms, ubiquitous network, ubiquitous computing, pervasive computing, and nomadic computer, are commonly mixed with one another. In the ubiquitous environment, users can be provided with any information they need anywhere anytime, and mobile service can be offered to mobile users. Therefore, the ubiquitous computing technology is emerging as a next-generation industry, and studies related to this are actively being carried out in the relevant academic world. Many of ubiquitous-related studies are researching on provision of more efficient services for users, which is because most users own mobile devices and are provided with various services, such as education, business, travel, and shopping, using the devices [3]. The growing significance of users in the ubiquitous environment is leading to lively studies of the ubiquitous decision support system (UDSS) for the purpose of supporting decision-making of users more efficiently.

2.2 Evaluation of UDS as a UDSS

As stated above, although there have been studies of ubiquitous steadily, studies of the UDSS have been insufficient relatively. This study intends to use the conceptual definitions that Kwon et al. [1] suggested of the UDSS.

Table 1. Competencies of UDSS and UDS level

Classification	Key competencies of UDSS	Level of UDS
Embeddedness	Wireless networking through compact intelligent device that is put inside of physical living space	Medium
Mobility	Working in the flexible mobile infrastructure of client devices	High
Nomadicity	Providing mobile users with computing and communication services that are transparent, integrated, and convenient	High
Portability	Offering a decision support system regardless of users' location using light and compact mobile devices	High

Since the study, one of the pioneering studies of the UDSS, deals with the scope, conceptual characteristics, and cases of the UDSS in detail, it is expected to be useful to understand this study. Kwon et al. [1] defines the UDSS as a system that supports users' decision-making process by enabling them to access to the intelligent space anywhere and anytime. The UDS, the subject of this study, can be assessed as follows at the level of the competencies of the UDSS that Kwon et al. [1] suggested. From the perspective of the UDSS, the UDS is not competent enough to offer information or services that are suitable for the characteristics of users using the self-intelligence.

Table 2. Examples of context-aware decision making using UDS

Decision-making process	Example
Intelligence	◦ John, a delivery man, wants to make decisions on delivery destination confirmation, delivery product check, customer check, delivery order, and one-day collection.
Design	◦ John, a delivery man, wants to know the best route to his delivery destinations for today. ◦ John, a delivery man, wants to know if he can visit his destination for one-day collecting from his current location. ◦ John, a delivery man, wants to know his best destination he can visit next in case his destinations are changed due to one-day collecting.
Choice	◦ John, a delivery man, can choose the best one among various alternatives

However, in terms of mobility, portability, and nomadicity, the UDS is providing various practical functions compared to services that simply use cell phones. Meanwhile, the interaction between the UDS and decision-making can be interpreted based on the decision-making process of Simon [4]. While the typical decision-making process is understood by the perceptual process of the decision makers, the decision-making process that is based on the connectivity and context-awareness function is defined by the decision-making circumstances that exist in problematic circumstances of the decision makers, for the decisions are made according to the environments and perceptions of the decision makers. Table 2 indicates examples of

decision-making using UDS that show the application examples of the UDSS in the ubiquitous environment.

3 Research Model and Hypotheses

Figure 1 indicates that the connectivity and context-awareness function of the UDS and perceived values have effects on the continuous intention to use through the decision satisfaction and trust. A hypothesis was built based on this conceptual model.

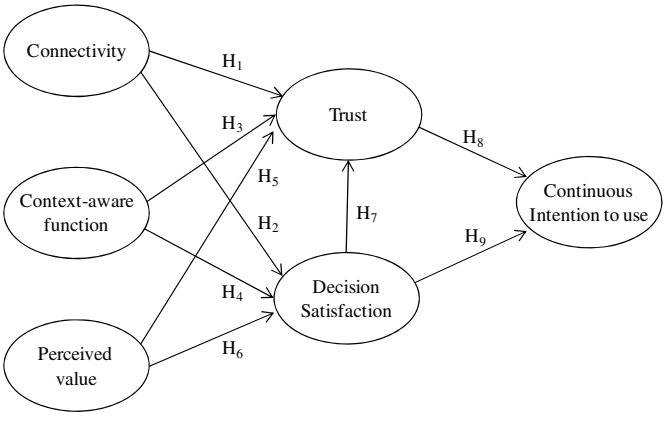


Fig. 1. Research Model

3.1 Connectivity

The concept of 'connectivity' that this study suggests encompasses both mobility and ubiquity. Therefore, the connectivity in the ubiquitous environment enables immediate access to customers, services, or information using the UDS, whenever needed and wherever the users are. Figge [5] also defines connectivity in terms of its accessibility to information regardless of time and location. In this respect, it can be interpreted that decision makers may be satisfied with the decision-making and trust it with confidence in its grounds when they can access to information or services they want to anywhere and anytime. Hence, we hypothesize:

Hypothesis 1: Connectivity has a positive effect on the trust.
Hypothesis 2: Connectivity has a positive effect on the decision satisfaction.

3.2 Context-Awareness Function

All people are surrounded by various forms, levels, and types of circumstances. Every 'moment of life', like when they feel need for something or when they want to purchase something, is included in circumstances. In other words, circumstances can be defines as the physical and mental state in which the user is in, comprehensively

considering time, place, and occasion surrounding the user. The context-awareness means the degree of the provision of the most effective information and services to users by considering overall circumstances in which users are, which is based on the localization and user identity of the ubiquitous environment [6]. The various information that the context-awareness function offers is expected to actively help the decision making process under various circumstances. This help will positively affect the decision satisfaction and trust. Hence, we hypothesize:

Hypothesis 3: Context-awareness function has a positive effect on the trust.
Hypothesis 4: Context-awareness function has a positive effect on the decision satisfaction.

3.3 Perceived Value

A perceived value in use of the UDS means a benefit or an advantage that the user receives as using the UDS. In marketing field, perceived values of users have been considered as one of the key variables that cause behavior of customers. The predominant view considers the perceived values of users as a trade off sacrifice and benefits. Putting together various views of researchers, a value can be defined as a benefit that is yielded as paying for products or services a company offers monetarily and non-monetarily in terms of a trade-off. In previous studies, perceived values of users are considered positive(+) when perceived quality is bigger than perceived sacrifice. This can be applied to the UDS; users of the UDS will not use the UDS unless the benefits of use of the UDS are bigger than the mental and temporal costs that they have paid to use a new system. If the users continue using the UDS, this generates decision satisfaction and trust of the UDS. Hence, we hypothesize:

Hypothesis 5: Perceived value has a positive effect on the trust.
Hypothesis 6: Perceived value has a positive effect on the decision satisfaction.

3.4 Decision Satisfaction and Trust

Decision satisfaction indicates whether users of the UDS are satisfied with the decision-making after using it. Since the key function of the UDS is to support the decision-making of users, it is very important to know whether the users have followed the decision support function of the UDS and been satisfied with it. Moreover, if the users of the UDS felt decision satisfaction are likely to trust and continuously use the UDS. The reason why the 'continuous' intention to use was asked instead of intention to use as a variable to measure the performance of the UDS is that the UDS is not really in its adoption stage in terms of diffusion stages of information systems. It is more crucial to estimate the intention to use the UDS continuously in the future because the UDS is already being used. If the users are satisfied with the decision support function of the UDS and eventually trust it, they will continuously use the UDS in the future. Hence, we hypothesize:

Hypothesis 7: Decision satisfaction has a positive effect on the trust.
Hypothesis 8: Trust has a positive effect on the continuous intention to use.
Hypothesis 9: Decision satisfaction has a positive effect on the continuous intention to use.

4 Research Methodology

4.1 Measurement Items

The variable of this study were measured according to Likert scale of 7 levels (1= strongly disagree, 7= strongly agree). The measuring items were modified and developed for this study, based on the measuring items that had been proved in their reliability and validity. As the circumstances of previous studies and those of this study were different, 30 respondents first were asked to answer the first draft of questions to verify the validity of the questions, which were changed afterwards into modified ones that were more acceptable to delivery men, considering the reaction of the respondents. Finally, the final questions were determined after discussions of delivery experts and relevant researchers.

4.2 Instrument Development and Data Collection

The main objective of this study is to examine the functions of the UDS as a form of the UDSS and empirically verify the intention to continuously use the UDS of delivery men who are using the UDS, considering the effects of connectivity, context-awareness function, and perceived values on it through decision satisfaction and trust. The objects that fulfilled all of these conditions were surveyed by D research institute, being interviewed for 20 to 25 minutes, and 5 dollars were rewarded to each of them. As a result, a total of 403 questionnaires were collected, but only 340 of them were analyzed after excluding those that had problems. All of the 340 respondents were men, and 244 of them (71.8%) were high school graduates. As for ages of the respondents, 92 of them (27.1%) were under 34 years old, 191 (56.2%) were between 34 and 44, 51 (15.0%) and 6 (1.8%) were over 56. In addition, as for the period of engagement in the delivery job, 191 of them (56.2%) were less than 3 years, 92 (27.1%) were between 3 and 6 years, 38 (11.2%) were 7 to 9 years, and 19 (5.6) of them were more than 10 years.

4.3 Measurement Model

Overall measurement quality is assessed using confirmatory factor analysis [7]. Although measurement quality is sometimes assessed factor by factor, in the current study each multiple-item indicator is considered simultaneously to provide for the fullest test of convergent and discriminant validity.

Table 3. Measurement model from confirmatory factor analysis[a]

Constructs and variables	Standardized factor loadings	CCR[a]	AVE[b]
Connectivity		0.902	0.698
1. I can access to the internet and obtain necessary services and information (ex. Delivery destination, delivery product, customer information, delivery information) whenever I want.	0.863		
2. I can use necessary services and information by accessing to the internet whenever I want, even when I am on the move.	0.863		
3. Interaction for obtaining necessary services and information is immediately available anywhere and anytime.	0.912		
4. The information and services are available when I need them no matter where I am.	0.793		
Context-awareness function		0.868	0.687
1. I can be provided with useful information (ex. Providing certain information at certain time to help delivery or customer visit).	0.826		
2. I can be provided with information and services that are suitable for the place I am in (ex. Conveying the one-day collection order in the area near from where I am delivering).	0.792		
3. I can be provided with useful information in accordance with my circumstances including the time and place.	0.867		
Perceived value		0.928	0.811
1. Using a scanphone reduces the cost for searching delivery destination.	0.887		
2. Using a scanphone reduces mental efforts required in the delivery processes and procedures.	0.792		
3. Using a scanphone reduces the general costs of the delivery process.	0.946		
Trust		0.838	0.635
1. The scanphone for delivery offers accurate information on customers, delivery destination, and delivery products.*	-		
2. The scanphone for delivery fulfills my needs.*	-		
3. The scanphone for delivery is reliable.	0.725		
4. The scanphone for delivery is related to my present and future benefits.	0.764		
5. Generally, I trust the scanphone for delivery.	0.892		
Decision Satisfaction		0.893	0.736
1. I am satisfied with the quality of the information and services that the scanphone for delivery provides.*	-		
2. I am satisfied with the quality of the system of the scanphone for delivery.	0.800		
3. I am satisfied with my use of the scanphone for delivery.	0.892		
4. I am generally satisfied with the scanphone for delivery.	0.879		
Continuous intention to use		0.957	0.881
1. I intend to use the PDA for delivery again.	0.932		
2. I intend to use the PDA for delivery as often as possible.	0.962		
3. I intend to use the PDA for delivery continuously in the future.	0.921		
4. I intend to recommend the PDA for delivery to others.*	-		

χ^2 = 356.681, d.f = 174, p = .000, GFI = 0.911, AGFI = 0.882, NFI = 0.935, CFI =0.965, RMSEA = 0.056.

[a] CCR : Composite Construct Reliability.

[b] AVE : Average Variance Extracted.

* Deleted items during Confirmatory Factor Analysis.

As shown in Table 3, four items with low factor loadings (below 0.50) are dropped from further analysis [7, 8]. All loadings exceed 0.50, and each indicator t-value exceeds 13.125 ($p < 0.001$). The χ^2 fit statistics show 244.327 with 137 degrees of freedom (χ^2/d.f = 1.783) ($p < 0.001$). The root mean square error of approximation (RMSEA) is 0.048; the comparative fit index (CFI) is 0.978; the adjusted goodness-of-fit index (AGFI) is 0.0.904; and the normed fit index (NFI) is 0.952. All statistics support the overall measurement quality given the number of indicators [7].

Furthermore, evidence of discriminant validity exists when the proportion of variance extracted in each construct exceeds the square of the Φ coefficients representing its correlation with other factors [9]. One pair of scales with a high correlation between them is perceived value and decision satisfaction ($\Phi = 0.544$, $\Phi^2 = 0.296$) (see Table 4). The variance extracted estimates are 0.811 and 0.736, indicating adequate discriminant validity. To allay concern about the discriminant validity of trust and continuous intention to use, the correlation between self presentation and community commitment is 0.488 ($\Phi^2 = 0.238$). The variance extracted estimates for these scales are 0.635 and 0.881, respectively. Thus, according to this assessment, the measures appear to have acceptable levels of validity.

Table 4. Descriptive statistics and correlations

Construct	Mean(S.D.)	(1)	(2)	(3)	(4)	(5)	(6)
(1) Connectivity	3.766(1.446)	1.000					
(2) Context-awareness	3.191(1.448)	0.419	1.000				
(3) Perceived Value	2.929(1.610)	0.349	0.335	1.000			
(4) Trust	3.639(1.248)	0.411	0.314	0.444	1.000		
(5) Decision satisfaction	3.022(1.397)	0.364	0.388	0.544	0.443	1.000	
(6) Continuous Intention	3.862(1.566)	0.335	0.358	0.350	0.488	0.457	1.000

4.4 Structural Model

Table 4 presents the maximum-likelihood estimates for the various overall fit parameters. The χ^2 statistic suggests that the data do not fit the model ($\chi^2 = 364.711$, df = 126, $p < 0.000$). However, because of the sensitivity of the χ^2 statistic to sample size it is not always an appropriate measure of the goodness-of-fit of the model. Therefore, multiple fit indices assess the overall evaluation of fit [10, 11]. The goodness-of-fit index (GFI) is 0.893; the Bentler and Bonett [12] normed fit index (NFI) is 0.889, respectively. Moreover, RMSEA is 0.075 and CFI is 0.924. These multiple indicators suggest that the model has good fit, justifying further interpretation.

Hypotheses H_1, H_3 and H_5 address the structural relationships among connectivity, context-awareness function, perceived value and trust. Connectivity has a positive effect on trust ($\beta = 0.201$, t-value = 3.133, $p < 0.01$), and is statistically significant at the $p < 0.01$ level, supporting H_1. But, context-awareness function has a no effect on trust ($\beta = 0.081$, t-value = 1.213, n.s). And, perceived value has a positive effect on trust ($\beta = 0.206$, t-value = 2.995, $p < 0.01$).

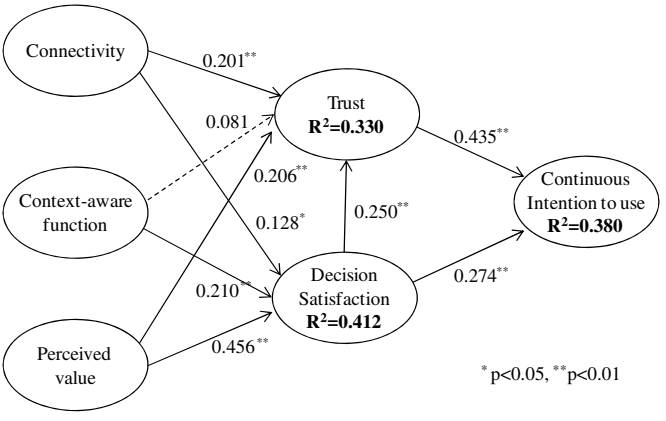


Fig 2. The estimated structural model

$\chi 2$ = 248.547, d.f = 140, p = 0.000, GFI = 0.930, AGFI=0.904, RMSEA = 0.048, NFI = 0.951, CFI = 0.978

Also, hypotheses H_2, H_4 and H_6 address the structural relationships among connectivity, context-awareness function, perceived value and decision satisfaction. Connectivity has a positive effect on decision satisfaction (β = 0.128, t-value = 2.198, $p < 0.05$), and is statistically significant at the $p < 0.05$ level, supporting H_2. H_4 is supported by the significant positive impact of context-awareness function on decision satisfaction (β = 0.210, t-value = 3.451, $p < 0.01$). Also, H_6 is supported by the significant positive impact of perceived value on decision satisfaction (β = 0.456, t-value = 7.783, $p < 0.01$). Decision satisfaction also has a positive impact on trust (β = 0.250, t-value = 3.387, $p < 0.01$), thus supporting H_7. Trust has a significant positive effect on continuous intention to use (β = 0.435, t-value = 6.901, $p < 0.001$), so H_8 is supported as well. Finally, decision satisfaction is associated with continuous intention to use (β = 0.274, t-value = 4.706, $p < 0.1$), supporting H_9.

5 Conclusions

This study empirically verified the impacts of the connectivity and context-awareness function, which are the key natures of the ubiquitous environment, on the decision-making process and the quality of the decisions of decision makers, considering the UDS used in the delivery field as a form of the UDSS. The results indicated that even in the current UDSS, which is at its early stage, the connectivity of ubiquitous aroused users' decision satisfaction and trust and had meaningful impacts on the continuous intention to use. It is thought to be a very significant result that the connectivity and context-awareness function, the key natures of the ubiquitous environment, are helpful for decision-making of decision makers who are actually using the UDSS to perform their work. It is expected that more meaningful and notable results will be drawn when the competencies of the UDSS become more diverse and personalized. The contributions of this study are that this study empirically verified the connectivity and context-awareness function of the ubiquitous environment by setting the UDS as a form of the early UDSS. Although previous studies have considered the

connectivity and context-awareness function, none of them reviewed the UDSS empirically by considering it within the ubiquitous environment. However, this study has the following limitations. The first limitation is whether the UDS, the study object, has fundamental functions of the UDSS sufficiently. To solve this problem, the researchers of this study reviewed various types of UDS in delivery fields and selected the objects that were the closest to the actual UDSS for analysis, but it still remains as a limitation. The second limitation is the frequency of use of the UDS of delivery men. According to the survey, there were some delivery men who thought that the UDS was making their jobs more complicated and made decisions according to personal experience. It was the limitation that this study could not relieve their bias. Future studies are required to examine if the UDS is arousing the intention of use and offering convenience and usefulness to individual decision makers.

Acknowledgment

This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Kwon, O., Yoo, K., Suh, E.: UbiDSS: a proactive intelligence decision support system as an expert system deploying ubiquitos computing technologies. Expert systems With Applications 28, 149–161 (2005)
2. Weiser, M.: The Computer for the 21st Century. Scientific American 265, 94–104 (1993)
3. Lyytinene, K., Yoo, Y.: Research Commentary: The Next Wave of Nomadic Computing. Information Systems Research 13, 377–388 (2002)
4. Simon, H.A.: Administrative Behaviour: A Study of Decision Making Processes in Administrative organization. Macmillan, New York (1957)
5. Figge, S.: Situation-dependent services-a challenge for mobile network operators. Journal of Business Research. Mobility and Markets: Emerging Outlines of M-Commerce 57, 1416–1422 (2004)
6. Kannan, P.K., Chang, A.-M., Whinston, A.B.: Wireless Commerce: Marketing Issues and Possibilites. In: Proceedings of the 34th Hawaii International Conference System Science. IEEE Computer Society, Los Alamitos (2001)
7. Anderson, J.C., Gerbing, D.W.: Assumptions and comparative strengths of the two-step approach. Sociological Methods & Research 20, 321–333 (1992)
8. Babin, B.J., Boles, J.S.: Employee behavior in a service environment: a model and test of potential differences between men and women. Journal of Marketing 62, 77–91 (1998)
9. Fornell, C., Larcker, D.F.: Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research 18, 39–50 (1981)
10. Bagozzi, R.P., Yi, Y.: On the Evaluation of Structural Equation Models. Journal of the Academy of Marketing Science 16, 74–94 (1988)
11. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L.: Multivariate data analysis, 6th edn. Prentice-Hall, Upper Saddle River (2006)
12. Bentler, P.M., Bonett, D.G.: Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. Psychological Bulletin 88, 588–606 (1980)

A Bayesian Network-Based Management of Individual Creativity: Emphasis on Sensitivity Analysis with TAN

Kun Chang Lee[1] and Do Young Choi[2,*]

[1] Professor of MIS at SKK Business School
WCU Professor of Creativity Science at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
kunchanglee@gmail.com
[2] Principal Consultant, LG CNS, Seoul 100-725, Republic of Korea
Tel.: +822-6363-5156; Fax: +822-6363-3905
dychoi96@gmail.com

Abstract. Creativity emerges as one of important resources for management. However, definitions of creativity have varied with researchers, and there is no universally agreed consensus about how to manage creativity in organizations. In this sense, managers who are interested in adopting specific type of creativity management strategy were confused. To avoid this problem, this study proposes a new method to creativity management by using a Bayesian Network (BN) that consists of nodes and arcs, and enables sensitivity analyses with various scenarios of interest. By focusing on individual creativity and its relationships with knowledge characteristic, intrinsic motivation, knowledge heterogeneity among team members, and organizational learning, we collected 222 valid questionnaires and performed what-if/goal seeking simulations based on TAN (Tree Augmented Naïve Bayesian Network) structure. Empirical results were promising and its practical meanings were well interpreted.

Keywords: Creativity management, Bayesian network, TAN, Sensitivity analysis, What-if analysis, Goal-seeking analysis.

1 Introduction

The importance of creativity as a source of individual and organizational performance has been emphasized. Moreover, recent organizations consider creativity as a strategic mean for enhancement of value creation in the competitive market environment. Studies of factors influencing creativity has been expanded from individual level based on the individual characteristic to group and organizational level based on social factors and contextual factors [12, 19]. As individuals conduct their tasks generally under circumstances of organization, various social and contextual factors should be considered simultaneously besides individual characteristic - intellectual capability, knowledge level, motivation, and self-efficacy, etc. - in order to explain creativity revelation processes [19]. This paper discusses the relationship between factors of individual characteristic and

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 512–521, 2011.

factors of organizational characteristic among factors influencing creativity. Furthermore, this paper is intended as an explanation of how these factors affect individual creativity. Even though various factors can be considered to explain individual creativity, we will consider two individual factors - intrinsic motivation and individual knowledge characteristic - and three organizational factors – organizational learning culture, knowledge sharing, and knowledge heterogeneity. In order to address these research questions, we will experiment the relationship among those factors influencing creativity with Bayesian network. In the traditional social researches, the regression method has been used to address these kinds of research questions. However, the tradition method has its rigorous functional form. For overcoming the rigidness of the regression, this paper uses a new approach with Bayesian Network. The Bayesian Network is unique method because the structure form of Bayesian Network explains the causality between target variable and all other variables. Also, it can help researchers interpret the results through experiments of what-if analysis and goal-seeking analysis. Therefore, we adopt the Bayesian Network to induce causality between six variables – individual creativity, individual knowledge characteristic, intrinsic motivation, knowledge sharing, organizational learning culture, and knowledge heterogeneity.

2 Literature Review

2.1 Creativity

Though creativity has been used variously in the different fields, it is considered as a series of processes that is capable of inducing innovative results based on the ability to create something new and innovative [2]. Studies of factors influencing creativity has been expanded from individual level based on the individual characteristics – personal characteristics, cognitive capabilities, and task environment - to group and organizational level, namely, group creativity and organizational creativity [12, 19]. Woodman et al. [19] insisted that various factors from individual level to group and organization level have an effect on creativity. Also they regarded that individual characteristics affecting creativity includes cognitive awareness, personality, intrinsic motivation, knowledge, etc. And there are various group characteristics and organizational characteristics affecting creativity such as cohesiveness, diversity, culture, resources, compensation, and structure. They found that these factors from three levels affect creative behavior and creative circumstances, and then creativity could be revealed through interactions. Therefore, when we address individual creativity revelation processes in organizations, we should consider organizational and contextual characteristics as well as personal characteristics. This paper focuses on motivation and individual knowledge among personal factors affecting creativity. There are two kinds of motivation – intrinsic motivation and extrinsic motivation. Intrinsic motivation has been treated as key factors by many researchers [3]. Intrinsic motivation refers to the motivational state in which an individual is attracted to their work in and of itself, not due to any external outcomes that might result from task engagement [5]. When people have higher intrinsic motivation, they have passion for their own tasks, and then they tend to contribute to their work [1]. Regarding social and organizational factors influencing on creativity, many researches emphasize on

organizational learning culture, knowledge sharing, and knowledge heterogeneity. Senge [16] addressed that learning organization means a continuous testing of experience and its transformation into knowledge available to whole organization and relevant to their mission. Garvin [8] defined learning organization as a competent organization that can create, acquire, and diffuse knowledge. Therefore, an organization with high level of organizational learning culture can positively affect creativity. Also knowledge sharing is related to creativity. Nelson and Cooprider [14] studied relationship between work performances and knowledge sharing. They found that knowledge sharing could enhance productivity as well as potential problem solving capability. Knowledge heterogeneity refers to variety of specialty that individuals possess [18]. When members hold different knowledge, skills, and capabilities, these kinds of heterogeneity can provide organizations useful perspectives and alternatives so that they enhance problem solving capabilities, organizational performances, and eventually creativity [15].

2.2 Bayesian Network

Recently complexities of decision making problems tend to grow exponentially as number of related variables increase, and causal relationships among them also get complicated much more. Creativity is no exception. Basically, creativity concept is elusive and abstract. Especially, from the perspective of managers and strategists, how to increase creativity on the either individual-level basis or team-level basis requires handling a large number of related factors in an organized way to come up with effective strategies of increasing the creativity. We argue that Bayesian network (BN) can be used as an extremely effective mechanism with which managers are able to locate strategic variables that seem to affect creativity significantly. Basically, BN is a graphical model that consists of a qualitative part(structural model) providing a visual representation of the interactions amid nodes(or variables) of interest, and a quantitative part(set of local probability distributions) allowing probabilistic inference and numerically measuring the impact of a node or a set of nodes on others [11]. Both the qualitative and quantitative parts determine a unique joint probability distribution over the variables in a specific problem [9]. In this sense, let us consider basic concepts of BN in a rather succinct way.

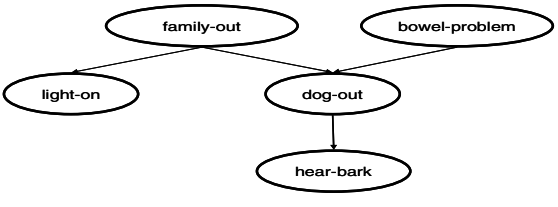


Fig. 1. Simple Bayesian Network (adapted from Charniak (1991))

First, BNs are directed acyclic graphs(DAGs) (like Figure 1), where the nodes are random variables, and thought of as variables(or factors) related to the target problem. The arcs specify the independence assumptions that must hold between the nodes. Such independence assumptions determine what probability information is required to

specify the probability distribution among the nodes in the BN. However, managers like to interpret the arcs as causality, which is right in some cases, and wrong in many cases. Therefore, how to interpret the arcs in BN requires caution. Second, BNs allow decision makers to calculate the conditional probabilities of the nodes in the network given that the values of some of the nodes have been observed. This is related to sensitivity analysis that this study focuses on. In Figure 1, for example, if you observe that the light is on(light-on = true) but do not hear your dog(hear-bark = false), the conditional probability of family-out given these pieces of evidence can be calculated(For this case, it is .5.). This process is about evaluation of the BN given the evidence. Then the conditional probability of the nodes given the changing evidence will also change our belief about the nodes. From this perspective, we believe that BN can be used very effectively for various kinds of sensitivity analyses with scenarios of interest within creativity managers' minds. Third, as noted previously, the arcs in BN can be interpreted as either causal relations or probabilistic relations. In Figure 1, the causal interpretation of the arcs says that the family being out has a direct causal connection to the dog being out, which, in turn, is directly connected to your hearing her. On the contrary, in the probabilistic interpretation, we adopt the independence assumptions that the causal interpretation suggests. Note that if you want to say that the location of the family was directly relevant to your hearing the dog, then you would have to put another arc directly between the two. From the structural point of view, the most general types of BNs are Naïve Bayesian Network(NBN) and Tree Augmented Naïve Bayesian Network(TAN). Though NBN has the simplest structure, in which a class node is linked with all of the children nodes, this study uses TAN [7] to overcome too rigid assumption of NBN such that independence between children variables should be maintained. In other words, to represent a wide variety of causal relationships existing among variables of interest in reality, we use TAN instead of NBN. In one word, TAN is an expansion of NBN into a tree shape.

3 Research Methodology and Experiment

3.1 Questionnaire Measurement and Survey

Survey research methods were typically used in the previous studies in order to reveal the relationships among factors affecting creativity. We adapted measurement items based on the reliable literature and conducted questionnaire survey on a 7-point, Likert-scale for this study. For example, in the case of individual creativity, we adapted four measurement items from the measures developed by Munoz-Doyague et al. [13] and Ettlie & O'Keefe [6]. To determine individual knowledge characteristic, three survey items were adapted from Hoegl et al. [10]. After constructing the questionnaire, 241 members were selected from the project teams in 14 software development companies and 12 system integration companies. Among responses, 222 responses were selected as suitable for this study.

3.2 Reliability and Confirmatory Factor Analysis

Six constructs were used in the questionnaire - individual creativity, individual knowledge characteristic, knowledge sharing, intrinsic motivation, knowledge

heterogeneity, and organizational learning culture. The confirmatory factor analysis was conducted for investigating validity and several measurement items were removed after principle component analysis with varimax method. As shown in the table 1, all constructs were considered to be reliable in the fact that the values of Cronbach's α of the six constructs were all greater than 0.6. The first factor, intrinsic motivation, explained 32.9% of the total variance. The total variance explained by the six factors was 71.5%.

Table 1. Reliability and Factor Analysis

Variables	Cronbach's alpha	MT	OL	KN	CR	HT	KS
MT3	0.909	**0.823**	-0.005	0.159	0.211	0.052	0.088
MT2		**0.814**	0.186	0.179	0.115	0.117	0.058
MT5		**0.810**	0.120	0.126	0.259	0.035	0.122
MT4		**0.810**	0.093	0.115	0.224	-0.039	0.155
MT1		**0.774**	0.148	0.123	0.105	0.293	-0.008
OL3	0.876	0.143	**0.821**	0.121	0.178	0.079	0.000
OL4		0.159	**0.816**	0.091	0.071	-0.016	0.010
OL5		-0.008	**0.790**	-0.015	0.186	0.040	0.039
OL1		0.046	**0.775**	0.050	0.045	-0.088	0.072
OL2		0.206	**0.710**	0.095	0.090	0.142	-0.032
OL7		0.011	**0.703**	-0.010	0.118	0.103	0.113
KN1	0.917	0.161	0.092	**0.887**	0.152	0.157	0.081
KN2		0.193	0.174	**0.858**	0.213	0.066	0.084
KN3		0.253	0.007	**0.834**	0.204	0.148	0.165
CR2	0.843	0.169	0.187	0.099	**0.762**	0.113	0.035
CR3		0.230	0.279	0.221	**0.723**	0.015	0.025
CR4		0.384	0.169	0.325	**0.690**	0.062	0.008
CR1		0.435	0.197	0.156	**0.608**	0.115	0.052
HT2	0.786	0.176	-0.007	0.019	0.057	**0.831**	0.208
HT1		-0.017	0.016	0.129	0.022	**0.800**	0.011
HT3		0.169	0.180	0.149	0.131	**0.794**	0.070
KS4	0.672	0.124	0.039	0.034	-0.001	0.096	**0.847**
KS3		0.048	0.153	0.099	-0.042	0.175	**0.807**
KS2		0.196	-0.045	0.314	0.329	-0.036	**0.543**
Eigenvalues		7.886	3.015	2.060	1.695	1.472	1.024
Variance explained (%)		32.9%	12.6%	8.6%	7.1%	6.1%	4.3%
Total variance explained (%)		32.9%	45.4%	54.0%	61.1%	67.2%	71.5%

Note: MT: Intrinsic Motivation, OL: Organizational Learning Culture, KN: Individual Knowledge Characteristic, CR: Individual Creativity, HT: Knowledge Heterogeneity, KS: Knowledge Sharing.

3.3 Experiment Results

The average values of each factor were calculated and then 7-point Likert-type scale was discretized into either Low(1 to 3), Middle(3 to 5), or High(5 to 7) in order to apply BN to the data. Basically, causal relationship among six variables and prediction accuracy were calculated by BN, as shown in Figure 2, where structure learning performed by WEKA (accessible from http://www.cs.waikato.ac.nz/ml/weka/) produces TAN as one of the best BN type for the empirical data. Regarding causal relationship,

all six variables have direct or indirect probabilistic relationship as shown in Figure 2. First, we found that all explanatory variables have direct relationship to target node, individual creativity. Second, intrinsic motivation is related to individual knowledge characteristic and knowledge heterogeneity. Likewise, individual knowledge characteristic is related to knowledge sharing and intrinsic motivation, and knowledge sharing is related to individual knowledge characteristic and organizational learning culture. The prediction accuracy of this model was 72.5% and causal relationships are easily identified using BN as shown in Figure 2.

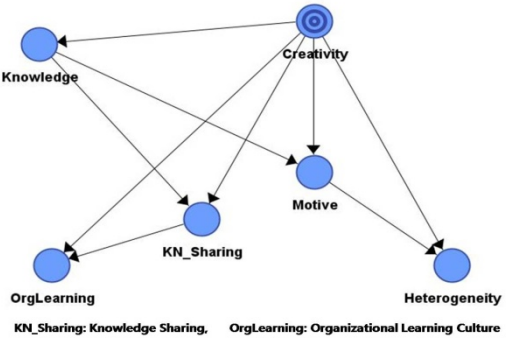


Fig. 2. Causal Relationship with individual creativity as a target node

In addition, the force of arc and node analysis was conducted in order to verify which node is relatively important and which relationship is relatively strong. As shown in the Figure 3, the most important node related to creativity is intrinsic motivation(global force of the node: 0.30), followed by individual knowledge characteristic(0.28), knowledge sharing(0.19), organizational learning culture(0.18), and knowledge heterogeneity(0.12). Also each node has its global contribution based on its own Kullback-Leibler(KL) divergence. Figure 3 shows the value of KL divergence and the value of global contribution for each node. From this analysis, we can find that the strongest relationship is the relation between intrinsic motivation and individual creativity(KL divergence: 0.14 and global contribution 17.8%), followed by the relation between individual knowledge characteristic and individual creativity(0.13, 16.9%), and organizational learning culture and individual creativity(0.11, 14.6%).

Scenario-based simulation(What-if and Goal-seeking analysis) was conducted based on two scenarios to find out how the posterior probabilities of each variable would change from the causal relationship in this model.

Scenario 1 (What-if analysis): Among explanatory variables, if the value of individual knowledge characteristic is high (i.e., it has a value between 5 and 7) and no other variables are changed, how do the individual creativity and other variables change?

As shown in Figure 4-(a), individual knowledge characteristic is related to individual creativity, intrinsic motivation, and knowledge sharing. Originally, the prior probability of middle individual creativity has the largest probability. However, if the

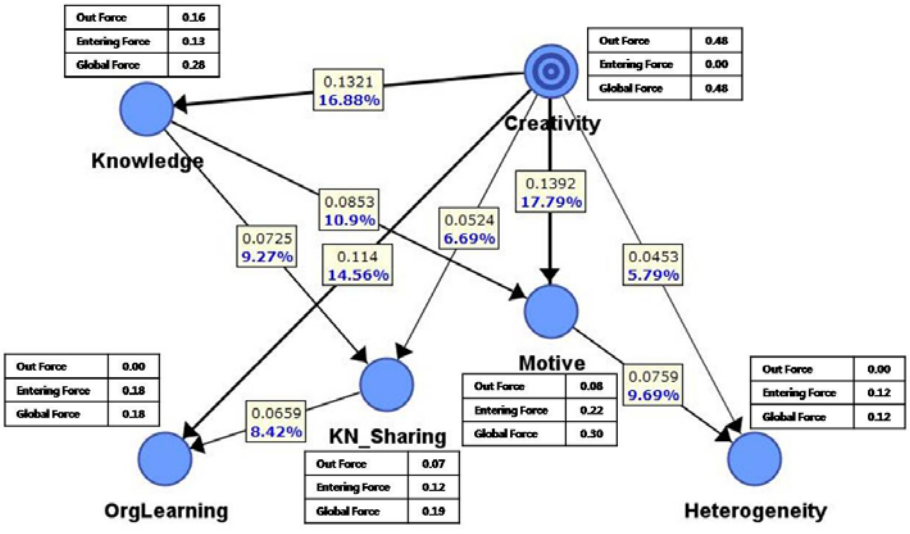


Fig. 3. Arc force and node force

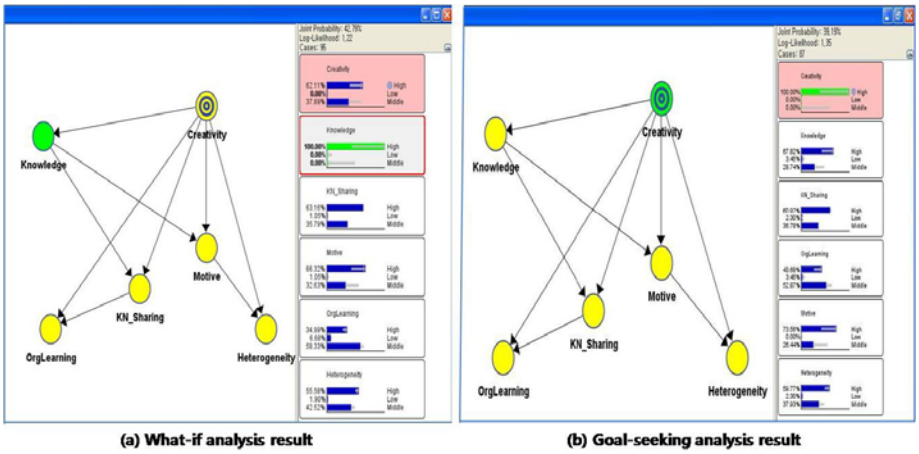


Fig. 4. Scenario Analysis Result

value of individual knowledge characteristic is set to be high, the posterior probability of the high individual creativity becomes largest. From this result, we can infer that increase in the value of individual knowledge characteristic helps increase the individual creativity level. Moreover, intrinsic motivation value and knowledge sharing value changes favorably most when the value of individual knowledge characteristic increases.

Scenario 2 (Goal-seeking analysis): For individual creativity to be high, what other factors should be changed?

As shown in the model above, individual creativity is affected by five variables: intrinsic motivation, individual knowledge characteristic, knowledge sharing, organizational learning culture, and knowledge heterogeneity. When the individual creativity is set to the highest level, the posterior probability for each high level of all of the variables increased. Among these variables, the levels of intrinsic motivation and those of individual knowledge characteristic increased when the value of individual creativity is set to be high as shown in Figure 4-(b). Also, other variables increased slightly if the individual creativity level increased.

4 Discussion

Through the BN based analysis with empirical data, it was found that those explanatory factors on which this paper focuses – individual knowledge characteristic, intrinsic motivation, knowledge sharing, organizational learning culture, and knowledge heterogeneity - directly affect individual creativity. In addition, we found the causality and relations among these variables. That is, intrinsic motivation is related to individual knowledge characteristic and knowledge heterogeneity. Likewise, individual knowledge characteristic has relationship with knowledge sharing and intrinsic motivation, and knowledge sharing is related to individual knowledge characteristic and organizational learning culture. Also it was found that what variables are relatively important and what causal relations are relatively stronger through the functions BN provides. The most important variable related to individual creativity seems intrinsic motivation, followed by individual knowledge characteristic, knowledge sharing, organizational learning culture, and knowledge heterogeneity. Also, the strongest relationship seems the relation between intrinsic motivation and individual creativity, followed by the relation between individual knowledge characteristic and individual creativity, and organizational learning culture and individual creativity. With these findings and several scenario-based simulation (what-if and goal-seeking analysis), we can have several implications for managers to conduct creativity strategically. First of all, managers should pay attention to intrinsic motivation enhancement and individual knowledge level enhancement of team members in order to increase creativity of the team as individual affecting factors. Second, BN results show that organizational learning culture affect knowledge sharing, and then affect individual knowledge and individual creativity. In addition, knowledge heterogeneity among team members expedites intrinsic motivation and individual creativity. Therefore, managers should encourage favorable organizational learning culture and consider maintaining diversity of team knowledge structure in terms of managing social affecting factors to creativity. Third, from the scenario-based simulation, the importance of intrinsic motivation and individual knowledge were reassured because they are the most influential factors to creativity. Therefore, intrinsic motivation and individual knowledge should be managed by organizations for strategic creativity management. Finally, managers can conduct more detailed simulations variously by scenarios referred to scenarios which were experimented in

this paper. We proved that Bayesian network is appropriate for analyzing the causality among related factors.

5 Conclusion

Recent organizations consider creativity as a strategic mean for enhancement of value creation in the competitive market environment. However, definitions of creativity have varied with researchers, and there is no universally agreed consensus about how to manage creativity in organizations. In this sense, managers who are interested in adopting specific type of creativity management strategy were confused. To avoid this problem, this study proposes a new method to creativity management by using a Bayesian Network that consists of nodes and arcs, and enables sensitivity analyses with various scenarios. Even though various factors can be considered to explain creativity, this paper considered two individual related factors - intrinsic motivation and individual knowledge characteristic - and three organizational related factors – organizational learning culture, knowledge sharing, and knowledge heterogeneity. We conducted experiments of the relationship among factors influencing individual creativity with Bayesian network. Through causality and scenario-based analysis of Bayesian network, we found that all factors – individual knowledge characteristic, intrinsic motivation, knowledge sharing, organizational learning culture, and knowledge heterogeneity - directly affect to individual creativity. Further, we found the causality and relations among these variables. Intrinsic motivation is related to individual knowledge characteristic and knowledge heterogeneity. Individual knowledge characteristic has relationship with knowledge sharing and intrinsic motivation, and knowledge sharing is related to individual knowledge characteristic and organizational learning culture. In addition, it was found that the relatively important variables were intrinsic motivation and individual knowledge characteristic. With the findings from causality and several scenario-based simulations by Bayesian network, we can have several important implications for managers. Managers should pay attention to intrinsic motivation enhancement and individual knowledge level enhancement of team members in order to increase creativity as individual factors. Further, managers should encourage favorable organizational learning culture and consider maintaining diversity of team knowledge structure in terms of managing social factors. Through this study, we proved that Bayesian network can be used as an extremely effective mechanism with which managers are able to locate strategic variables that seem to affect creativity significantly. Nevertheless, there are some considerations for further study. For example, we considered only five variables for explaining creativity. Other factors should be considered such as social network structure for rich understanding of team level and organizational level creativity.

Acknowledgment

This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Amabile, T.M.: The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology 45(2), 357–376 (1983)
2. Amabile, T.M.: A model of creativity and innovation in organizations. Research in Organizational Behavior 10, 123–167 (1988)
3. Barron, F.B., Harrington, D.M.: Creativity, intelligence, and personality. Annual Review of Psychology 32, 439–476 (1981)
4. Charniak, E.: Bayesian networks withoug tears. AI Magazine, 50–63 (1991)
5. Deci, E.L., Ryan, R.M.: Intrinsic motivation and self determination. Social Networks 1, 215–239 (1979)
6. Ettlie, J.E., O'Keefe, R.D.: Innovative attitudes, values, and intentions in organizations. Journal of Management Studies 19(2) (1982)
7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29(2), 131–163 (1997)
8. Garvin, D.A.: Building a learning organization. Harvard Business Review 71(4), 78–91 (1993)
9. Cooper, G.F.: An overview of the representation and discovery of causal relationships using Bayesian networks. In: Glymour, C., Cooper, G.F. (eds.) Computation, Causation & Discovery, pp. 3–62. AAAI Press/MIT Press, Cambridge, MA (1999)
10. Hoegl, M., Parboteeah, K.P., Munson, C.L.: Team-level antecedents of individuals' knowledge networks. Decision Sciences 34(4) (2003)
11. Jensen, F.V.: Bayesian networks. WIREs Computational Statistics 1(1), 307–315 (2009)
12. Kurtzberg, T.R., Amabile, T.M.: From Guilford to creative synergy: opening the balck box of team-level creativity. Creativity Research Journal 13, 285–294 (2001)
13. Munoz-Doyague, M.F., Gonzalez-Alvarez, N., Nieto, M.: An examination of individual factors and employees' creativity: The case of Spain. Creativity Research Journal 20(1), 21–33 (2008)
14. Nelson, K.M., Cooprider, J.G.: The contribution of shared knowledge to IS group performance. MIS Quarterly 20(4), 409–432 (1996)
15. Pelled, L.H., Eisenhardt, K.M., Xin, K.R.: Exploring the black box: An analysis of work group diversity, conflict, and performance. Administrative Science Quarterly 44(1), 1–28 (1999)
16. Senge, P.M.: The fifth discipline: The art and practice of the learning organization. Doubleday, New York (1990)
17. Spirtes, P., Glymour, C., Scheines, R.: Causation prediction and search. In: Berger, et al. (eds.) Lecture Notes in Statistics, 1st edn., vol. 81. Springer, Berlin (1993)
18. Tiwana, A., McLean, E.R.: Expertise integration and creativity in information systems development. Journal of Management Information Systems 22(1), 13–43 (2005)
19. Woodman, R.W., Sawyer, J.E., Griffin, R.W.: Toward a theory of organizational creativity. Academy of Management Review 18(2), 293–321 (1993)

General Bayesian Network Approach to Balancing Exploration and Exploitation to Maintain Individual Creativity in Organization

Kun Chang Lee[1] and Min Hee Hahn[2,*]

[1] Professor of MIS at SKK Business School
WCU Professor of Creativity Science at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
kunchanglee@gmail.com
[2] Researcher, Business Management Unit, LG CNS CO., Ltd.
Seoul 100-725, Republic of Korea
Tel.: +82 2-6363-5184, Fax: +82 2-6363-3300
minheehahn@gmail.com

Abstract. As market competition grows fierce, how to maintain competitiveness in the market emerges a crucial issue for all the organizations. Aware of this urgent fact, companies have been seeking best way of managing their creativity at competitive levels. However, most of existing approaches currently discussed in literature were limited to narrative and elusive statements from which creativity management strategists could not extract set of concrete action rules. To overcome this pitfall, this study proposes a novel approach to creativity management by adopting General Bayesian Network (GBN). Especially, based on the findings from literature that balance of exploration and exploitation leads to sustainable management of creativity, we built a research model including the five variables affecting individual creativity such as exploration, exploitation, task complexity, bureaucratic culture, and supportive culture. To induce a set of causal relationships among the individual creativity and the five variables, GBN was applied to 227 valid questionnaire sample data. Through the what-if and goal-seeking simulations, promising empirical results were obtained, which shed robust and meaningful platform for further studies in this exciting field.

Keywords: Individual creativity, Exploration, Exploitation, Organizational culture, Task complexity, General Bayesian Network.

1 Introduction

From the several decades ago, companies have consumed much of their conventional management resources to build and exercise strategies more effectively in the competitive market. Problems with this approach are that there remain few tangible resources they can spend further, and intangible resources are necessary to be found and applied to enhancing the effectiveness of management strategies. In this respect, creativity

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 522–531, 2011.

management has emerged as an important strategy. However, existing studies about creativity were limited to just discussing academic issues, most of which practitioners cannot adopt in reality. To overcome this pitfall, this study proposes a new approach to individual creativity management by using a General Bayesian Network (GBN).

It is widely known that creativity has a wide variety of variables that may affect it, ranging from organizational culture, leadership to individual knowledge level [39]. However, for the practitioners to apply academic findings to the real issue of creativity management, extraction of causal relationships from among the set of creativity and relevant variables is necessary. GBN [9, 20] seems appropriate from this perspective, because GBN has been successfully applied to resolving highly complicated decision making problems [8, 9, 20]. By applying the GBN to survey data, we successfully extracted causal relationships between exploration, exploitation, and other relevant factors affecting individual creativity.

Empirical findings revealed that from taking advantage of the flexible structure and inference capabilities supported by the GBN, balancing exploration and exploitation can be obtained very effectively by adjusting related factors such as task complexity, bureaucratic culture, and supportive culture.

2 Previous Studies

2.1 Individual Creativity

From the existing literature, we confirmed that the concept of creativity has expanded into diversified fields, including the arts, science, and business disciplines (e.g., [35, 36, 38]). Creativity is complex concept that researchers defined in different ways [33]. Therefore, there are many definitions about creativity. Typically, Amabile [3] defined creativity as the "production of novel and useful ideas". In other words, creativity can be defined as any process used to generate creative outcomes based on the ability to produce something new [3]. This definition has been cited in later conceptual models [39] and in various studies [2, 27].

Guilford [11] argued that creativity is a continuous trait in all people and that individuals with recognized creative talent simply have "more of what all of us have." After Guilford's study, researchers have mainly centered on "individual" creativity. For example, Amabile [4, 2] suggested that creativity has a greater chance to occur when people's creative abilities and expertise overlap with their strongest intrinsic interests and that the greater the level of each one of the components the greater will be the creativity. This is called the ''creativity intersection'' [3], so that an individual possessing the three components (i.e., domain-relevant knowledge, creativity relevant skills, and intrinsic motivation) will have a higher probability for being creative. Besides, the levels of each one of his joint components determine the final level of creativity reached by an individual.

2.2 Exploitation and Exploration

The concept of exploration and exploitation introduced by March [21] as follows: "Exploration includes things captured by terms such as search, variation, risk taking, experimentation, flexibility, discovery, and innovation. Exploitation includes such

things as refinement, choice, production, efficiency, selection, implementation, and execution." Moreover, Levinthal and March [17] added a definition that exploration is related to a pursuit of new knowledge, whereas exploitation is related to the use and development of things already known. After them, many researchers have had various distinctions between exploration and exploitation (e.g., [5, 12, 16]).

Many researchers have been interested in exploration and exploitation as independent variables for corporate performance. For instance, Rosenkopf & Nerkar [29] investigated the domain and overall impact of exploration on technological evolution within or across organizational or technological boundaries and distinguished between different types of explorations. Nerkar [26] investigated impacts of temporal exploitation and exploration on later knowledge creation. And Ahuja & Lampert [1] examined the impact of exploratory strategies on the number of breakthrough inventions by a firm in the subsequent period.

On the other hand, there are many studies considering exploration and exploitation as dependent variables. For example, Benner & Tushman [6] studies the influence of process management on exploitative and explorative innovations.

In these ways, the studies about exploration and exploitation have used the two constructs as dependent variables, and examined the influence of special independent variables on them or the effect of diverse independent variables on firm's performance and its substitutes. In this study, we don't try to put out comprehensive interpretations about these, but we attempt to approach viewpoint of ambidexterity exploration-exploitation, based on existed studied references.

2.3 Organizational Culture

Although the universal definition of organizational culture has been elusive [18], it is defined as patterns of values, beliefs and assumptions shared by members within organization [34, 31]. The values, beliefs, and assumptions underlying an organization's culture bind its employees together and become the manner or strategies through which the organization achieves its goals [22].

From literatures, there have been the diversified influences of organizational culture on individuals and organization. For example, organizational culture influenced not only knowledge circulation and its sharing, but is an important factor in knowledge management field indispensible with e-business and knowledge management in organization [19]. Moreover, organizational culture has also been shown to impact on ways for members to operating units and playing their roles in many facets [24]. Martins and Terblanche [23] regarded it as a critical component of organization's success. Besides, Ruppel and Harrington [30] discovered that organizational culture has influence on Intranet's implementation and found that the positive effects of organizational culture, emphasizing 'care, flexibility and innovative policies and procedures', on knowledge circulation and dissemination are expected.

Organizational culture in this paper is based on studies of Hill and Jones [13], Wallach [37] and Lin [19]. Hill and Jones [13] defined the organizational culture as the "common value and rule specified for internal members." Organizational culture is grouped into two types [37] : (1) Bureaucratic culture: means that most of the work in an organization is standardized and operates on the basis of control and power. Tasks are completed in proper sequence and enterprise ethic is specially emphasized. (2)

Supportive culture: means an open and harmonious working environment. Participation, teamwork and interpersonal relationship are specially emphasized.

2.4 Bayesian Network

Bayesian Network (BN) is a powerful formalism for representing a joint probability distribution on a set of statistical variables. It can represent domain knowledge and its uncertainties, making possible reasoning with uncertainties. Basically, Bayesian networks are directed acyclic graphs (DAG) that allow for efficient and effective representation of joint probability distributions over a set of random variables. Formally, a BN consists of two parts. The first part is a DAG consisting of nodes and arcs. The nodes are the variables X_1,X_2, . . . , X_n in the data set whereas the arcs indicate direct dependencies between the variables. The second part represents the conditional probability distributions.

The BN can be used as a classifier when users want to determine whether the exact probability of an event is above (or below) a certain threshold [7]. When the BN is used as a classifier, a class node should be designated in advance. The class node then becomes a target node with which other nodes are interlinked depending on the structure.

To improve classification performance, Cheng and Greiner [9] suggested two BN structures called a Bayesian Network Augmented NBN (BAN) and a General Bayesian Network (GBN). A BAN allows all other nodes to be direct children of the class node, but a complete BN is constructed between the child nodes. Meanwhile, a GBN is a full-fledged BN in which causal relationships between the class node and all other nodes are flexibly formulated using an efficient network construction technique based on conditional independence tests [8].

3 Research Methodology and Experiment

3.1 Questionnaire Survey and Sample Statistics

To analyze the relationships among factors affecting individual creativity, survey items were adapted from previous studies. For example, formed from 4 items obtained from Ettlie and O'Keefe [10], Scott and Bruce [32], Zhou and George [40], and Munoz-Doyague et. al. [25], a 7-point scale was made in order to measure individual creativity. The scale measured the concept's two fundamental aspects, namely novelty and utility. The items were averaged out and a high score represented an employee with a highly creative value. And based on Lee and Choi [15], Katila and Ahuja [14], and Prieto et. al. [28], exploitation and exploration have been measured by using 9 items, four items concerning to exploration and five items concerning to exploitation. The first four items measured the degree to which the business proposal development introduce new ideas, new knowledge and cover and correct problems areas where customers were unsatisfied. The last five items measured the degree to which the business proposal development introduces lessons learnt in the past, existing competences, and combines and integrate different knowledge.

After constructing the questionnaire, the survey was conducted people for who is working in IT companies in South Korea and has experience at the proposal project.

Among the initial response of 240 persons composed of 44 teams, 227 responses are included in the survey target. Among them, there are 194 males and 33 females. In terms of age group, 29 are in the age 20 to 30, 122 are in the age 30 to 40, the highest among the groups, 67 are in the age 40 to 50, and 9 are in the age 50 or higher.

3.2 Reliability and Confirmatory Factor Analysis

Six constructs were used in the research question, namely, task complexity, organizational culture(i.e., bureaucratic culture and supportive culture), exploitation, exploration, and individual creativity. In Table 1, Cronbach's alpha values for the six constructs were all greater than 0.7, indicating that these items were reliable. In addition, principle component analysis with the verimax rotation option was used to test the validity of each item. The first factor, exploitation, explained 32.3% of the total variance; the second factor, individual creativity, 9.7%. The total variance explained by the six factors was 66.5%. On the basis of these results, we concluded that the questionnaire items were statistically valid.

Table 1. Reliability and Factor analysis

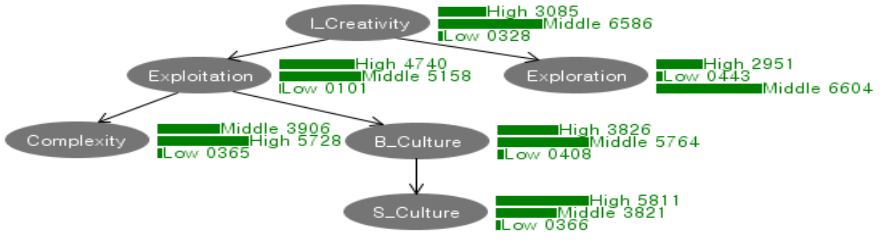
Variables	Cronbach's alpha	ET	IC	TC	ER	SP	BR
ET4	0.81	**0.72**	0.07	0.18	0.05	0.08	0.24
ET2		**0.70**	0.11	0.10	0.36	0.08	0.04
ET5		**0.66**	0.21	0.24	0.11	0.24	-0.01
ET1		**0.66**	0.22	0.07	0.35	0.16	0.06
ET3		**0.64**	0.23	0.03	-0.01	0.09	0.11
IC1	0.85	0.09	**0.77**	0.08	0.22	0.18	0.05
IC4		0.27	**0.76**	0.17	0.29	0.09	0.04
IC3		0.16	**0.75**	0.08	0.26	0.08	0.18
IC2		0.31	**0.70**	0.07	0.08	0.03	0.16
TC2	0.81	0.11	0.01	**0.80**	0.07	0.07	0.10
TC3		0.16	-0.01	**0.78**	0.21	0.07	0.12
TC4		0.09	0.17	**0.74**	0.07	0.13	0.18
TC1		0.11	0.17	**0.74**	-0.11	0.14	-0.04
ER3	0.79	-0.06	0.26	0.00	**0.80**	0.02	0.24
ER2		0.23	0.14	0.10	**0.69**	0.10	0.08
ER1		0.37	0.23	0.09	**0.65**	0.08	-0.01
ER4		0.15	0.20	0.05	**0.62**	0.24	0.14
SP3	0.79	0.09	0.17	0.24	0.13	**0.82**	0.04
SP1		0.17	0.11	-0.01	0.02	**0.78**	0.21
SP2		0.17	0.02	0.21	0.21	**0.76**	-0.03
BR2	0.76	0.18	0.13	0.20	0.09	0.07	**0.82**
BR1		0.05	0.14	0.08	0.22	0.02	**0.80**
BR3		0.22	0.09	0.09	0.10	0.48	**0.61**
Eigenvalues		7.43	2.23	1.58	1.52	1.35	1.20
Variance explained (%)		32.3	9.7	6.9	6.6	5.9	5.2
Total variance explained (%)		32.3	42.0	48.8	55.5	61.3	66.5

* Note : ET: Exploitation, IC: Individual creativity, TC: Task Complexity, ER: Exploration, SP: Supportive culture, BR: Bureaucratic culture.

3.3 Results

To apply the GBN mechanism to the questionnaire data, we calculated the average values for each factor and then transformed the Likert scale for each factor into either Low(1 to 3), Middle(3 to 5), or High(5 to 7). Causal relationships among the seven variables were depicted by the GBN results, as shown in Figure 1.

First, the two variables of exploitation and exploration were found to have a direct relationship to individual creativity. Second, exploitation is related to task complexity and bureaucratic culture. Likewise, bureaucratic culture is associated with supportive culture. The prediction accuracy of this model was 74.01 %. Consequently, casual relationships are able to be easily identified using GBN.



* Note : I_Creativity: Individual creativity, Complexity: Task Complexity, S_Culture: Supportive culture, B_Culture: Bureaucratic culture

Fig. 1. GBN with individual creativity as a target node

In order to know the what-if and goal-seeking support capability of the GBN, let us consider the following two scenarios, in which sensitivity analysis is performed by taking advantage of causal relationships suggested by Figure 1.

Scenario 1 (What-if Analysis) : *With ambidexterity perspective, among five values, if the exploitation and exploration are high at the same time (i.e., each of them has a value between 5 and 7) and no other variables are changed, how do the individual creativity and other variables change?*

As shown in Figure 2, exploitation and exploration are related with individual creativity, task complexity, and bureaucratic culture. Originally, the prior probability of middle individual creativity has the largest probability. However, when the exploitation and exploration are set to be high, the posterior probability of the high individual creativity becomes largest. This means that increase in exploitation and exploration help increase the individual creativity level. In addition, bureaucratic culture changes favorably most when exploitation and exploration value increases.

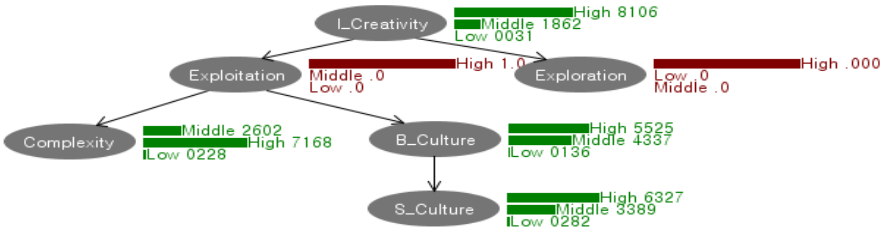


Fig. 2. What-if analysis result

Scenario 2 (Goal-Seeking analysis) : *For Individual creativity to be high, what other factors should be changed?*

Consequently, the individual creativity is affected mainly by exploration. In other words, for individual creativity to be high, exploration factor should be changed. In addition, when the individual creativity is set to the highest level, the posterior probability for each high level of all of the variables increased. Notably, when the individual creativity is set to the highest level, posterior probability changing of the supportive culture is greater than the bureaucratic culture. Eventually, to derive individual creativity of high level, it is showing that needs to construct organizational supportive culture.

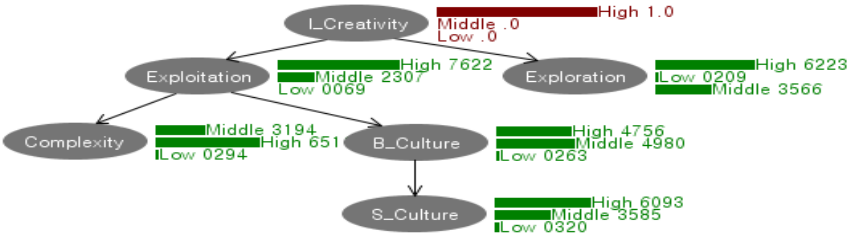


Fig. 3. Goal-seeking analysis result

3.4 Discussion

We implemented a Bayesian network simulation for the members of proposal teams in SI companies to analyze the individual creativity. From analyzing the GBN results, it was found that those factors mainly affecting the individual creativity are exploitation and exploration. In addition, exploitation is related to task complexity and bureaucratic culture. Bureaucratic culture is associated with supportive culture. And we found that the most important node affecting the individual creativity is the exploration. Moreover, we verified, if both exploitation and exploration are increasing at the same time, maximize the individual creativity. There are several implications for the IT companies.

First of all, both exploitation and exploration are important factors for individual creativity in SI proposals. According to the level of exploitation and exploration, significant differences existed with regard to influence on individual creativity. Therefore, IT corporations should be recognized that both exploitation and exploration are critical antecedents of individual creativity.

Second, GBN results tell that the exploitation is related to task complexity and bureaucratic culture. And bureaucratic culture is associated with supportive culture. Exploitation is utilization and development of existing knowledge. Thus, active exploitation in order to make this happen, rather than too easy task, relatively high levels of complexity task. And companies build an appropriate level of control based on supportive culture is also needed.

Finally, we proved that GBN is suitable for analyzing the causal relationship among related factors. In addition, by what-if and goal-seeking analysis, more detailed analysis which is difficult in traditional methods like regression analysis is possible.

4 Concluding Remarks

We proposed using the GBN to obtain a set of causal relationships and build useful management strategies about individual creativity by performing a variety of sensitivity analyses.

Findings from using the GBN are as follows. First, we found that individual creativity is closely related to exploitation and exploration. Second, exploitation is associated with task complexity and bureaucratic culture. Bureaucratic culture is related to supportive culture. Third, through goal-seeking analysis, in order to increase individual creativity, you should enhance the exploration. But the effect is greater when enhance both exploitation and exploration at the same time.

Through our experience with using the GBN, we came to firmly believe that GBN has great potentials enough to provide causal relationships as well as sensitivity analyses capability, all of which could be extremely useful for IT companies to understand which factors are probably crucial in determining individual creativity. Derivation of the GBN-based individual creativity enhancement strategies like this would lead to mutually beneficial outcomes for both IT companies and customers.

Nevertheless, future study issues still remain unanswered. First, a number of control variables including gender, age, and other individual related factors need to be introduced. Second, our study targeted members of a proposal team who worked together for only a short time period. Therefore, our results may be difficult to generalize to entire organizations. It is found that the study needs to generalize its results by allowing the analysis pool to include other groups besides the proposal teams and conduct comparison study between the more diversified groups in a bigger pool. Finally, the structural equation modeling approach needs to be applied to induce statistically significant results.

Acknowledgment. This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Ahuja, G., Lampert, J.: Entrepreneurship in the Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Inventions. Strategic Management Journal 22, 521–543 (2001)

2. Amabile, T.M.: Creativity in context. Westview Press, Boulder.CO (1996)
3. Amabile, T.M.: A Model of Creativity and Innovation in Organizations. Research in Organizational Behavior 10, 123–167 (1988)
4. Amabile, T.M.: The social psycology of creativity: A componential conceptualization. Journal of Personality and Social Psycology 45, 357–376 (1983)
5. Beckman, C., Haunschild, P.R., Phillips, D.: Friends or Strangers? Firm-Specific Uncertainty, Market Uncertainty, and Network Partner Selection. Organization Science 15, 259–275 (2004)
6. Benner, M.J., Tushman, M.L.: Process management and technological innovation: A longitudinal study of the photography and paint industries. Administrative Science Quarterly 47, 676–706 (2002)
7. Chan, H., Darwiche, A.: Reasoning about Bayesian Network Classifiers. In: Meek, C., Kjærulff, U. (eds.) 19th Conference in Uncertainty in Artificial Intelligence, pp. 107–115 (2003)
8. Cheng, J., Bell, D.A., Liu, W.: Learning Belief Networks from Data: An Information Theory Based Approach. In: 6th ACM International Conference on Information and Knowledge Management, pp. 325–331 (1997)
9. Cheng, J., Greiner, R.: Learning Bayesian Belief Network Classifiers: Algorithms and System. In: 14th Canadian Conference on Artificial Intelligence, pp. 141–151 (2001)
10. Ettlie, J.E., O'Keefe, R.D.: Innovative attitudes, values, and intentions in organizations. Journal of Management Studies 19, 163–182 (1982)
11. Guilford, J.P.: Creativity. American Psychologist 5, 444–454 (1950)
12. He, Z.L., Wong, P.K.: Exploration vs Exploitation: An Empirical Test of the Ambidexterity Hypothesis. Organization Science 15, 481–494 (2004)
13. Hill, C.W.L., Jones, G.R.: Strategic management – an integrated approach, 4th edn., Boston, HM (1998)
14. Katila, R., Ahuja, G.: Something old, something new: a longitudinal study of search behaviour and new product introduction. Academy of Management Journal 45, 1183–1194 (2002)
15. Lee, H., Choi, B.: Knowledge management enablers, processes, and organizational performance: an integrative view and empirical examination. Journal of Management Information Systems 20, 179–228 (2003)
16. Lee, J., Lee, J., Lee, H.: Exploration and Exploitation in the Presence of Network Externalities. Management Science 49, 553–570 (2003)
17. Levinthal, D.A., March, J.G.: The myopia of learning. Strategic Management Journal 14, 95–112 (1993)
18. Lewis, D.: Five years on—The organizational culture saga revisited. Leadership & Organization Development Journal 23, 280–287 (2002)
19. Lin, W.: The Effect of Knowledge Sharing Model. Expert Systems with Applications 34, 1508–1521 (2008)
20. Madden, M.G.: On the Classification Performance of TAN and General Bayesian Networks. Knowledge-Based Systems 22, 489–495 (2009)
21. March, J.G.: Exploration and exploitation in organizational learning. Organization Science 2, 71–87 (1991)
22. Marcoulides, G.A., Heck, R.H.: Organizational culture and performance: proposing and testing a model. Organization Sciences 4, 209–225 (1993)
23. Martins, E.G., Terblance, F.: Building organisational culture that stimulates creativity and innovation. European Journal of Innovation Management 6, 64–74 (2003)

24. McDermott, C.M., Stock, G.N.: Organizational culture and advanced manufacturing technology implementation. Journal of Operations Management 17, 521–533 (1999)
25. Munoz-Doyague, M.F., Gonzalez-Alvarez, N., Nieto, M.: An Examination of Individual Factors and Employees' Creativity: The Case of Spain. Creativity Research Journal 20, 21–33 (2008)
26. Nerkar, A.: Old is good? The value of temporal exploration in the creation of new knowledge. Management Science 49, 211–229 (2003)
27. Oldham, G.R., Cummings, A.: Employee creativity: Personal and contextual factors at work. Academy of Management Journal 39, 607–634 (1996)
28. Prieto, I.M., Revilla, E., Rodriguez-Prado, E.: Managing the knowledge paradox in product development. Journal of Knowledge Management 13, 157–170 (2009)
29. Rosenkopf, L., Nerkar, A.: Beyond Local Search: Boundary-Spanning Exploration, and Impact in the Optical Disk Industry. Strategic Management Journal 22, 287–306 (2001)
30. Ruppel, C.P., Harrington, S.J.: Sharing knowledge through intranets: a study of organizational culture and intranet implementation. IEEE Transactions on Professional Communication 44, 37–52 (2001)
31. Schein, E.H.: Organizational Culture and Leadership: A Dynamic View. Jossey-Bass Publishers, San Francisco (1985)
32. Scott, S.G., Bruce, R.A.: Determinants of innovative behaviour: A path model of individual innovation in the work place. Academy of Management Journal 37, 580–607 (1994)
33. Shalley, C.E., Gilson, L., Blum, T.C.: Matching creativity requirements and the work environment: Effects on satisfaction and intentions to leave. Academy of Management Journal 43, 215–223 (2000)
34. Sigler, T., Pearson, C.: Creating and empowering culture: examining the relationship between organizational culture and perceptions of empowerment. Journal of Quality Management 5, 27–52 (2000)
35. Stumpf, H.: Scientific creativity: A short overview. Educational Psychology Review 7, 225–241 (1995)
36. Tang, P.C., Leonard, A.R.: Creativity in art and science. Journal of Aesthetic Education 19, 5–19 (1985)
37. Wallach, E.J.: Individuals and organizations: the cultural match. Training and Development Journal 37, 45–76 (1983)
38. Williams, W.M., Yang, L.T.: Organizational creativity. In: Sternberg, R.J. (ed.) Handbook of Creativity, pp. 373–391. Cambridge University Press, Cambridge (1999)
39. Woodman, R.W., Sawyer, J.E., Griffin, R.W.: Toward a theory of organizational creativity. Academy of Management Review 18, 293–321 (1993)
40. Zhou, J., George, J.M.: When job dissatisfaction leads to creativity: Encouraging the expression of voice. Academy of Management Journal 44, 682–697 (2001)

The Role of Cognitive Map on Influencing Decision Makers' Semantic and Syntactic Comprehension, and Inferential Problem Solving Performance

Soon Jae Kwon[1], Kun Chang Lee[2,*], and Emy Elyanee Mustapha[3]

[1] Professor of MIS, Department of Business Administration
Daegu University, Kyong San 712-714, Republic of Korea
kwonsj72@gmail.com
[2] Professor at SKK Business School and WCU Professor at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic Korea
Tel.: 82 2 7600505; Fax: 82 2 7600440
kunchanglee@gmail.com
[3] Doctoral Candidate, Department of Business Administration
Daegu University, Kyong San 712-714, Republic of Korea

Abstract. In the field of decision making, cognitive map (CM) has been successfully applied to resolving a wide variety of complicated decision making problems. However, in literature, it is very rare to find those studies investigating the influence of CM on decision maker's semantic and syntactic comprehension, and inferential problem-solving performance. To pursue this research issue, we suggest the empirical findings from the rigorous experiment where participants were invited from those having experience with six sigma projects for years. To systematically test the effect of CM, participant were grouped into two expertise types (experts vs novice) and two types of CM method knowledge (high CM knowledge vs low CM knowledge). Experimental results showed that CM can be used in significantly enhancing decision makers' semantic and syntactic comprehension, as well as inferential problem-solving.

Keywords: Cognitive map, Semantic comprehension, Syntactic comprehension, Inferential problem-solving, Six sigma.

1 Introduction

Causal map (CM) is used to capture perception of decision makers (DMs) faced with complex and unstructured decision problems. Many relevant literatures showed that a CM can be used for solving many kinds of decision problems [1], [2], [3], [4], [5], [6], [7], [8], [9], most of which belong to unstructured decision problems. And, CM can describe and facilitate elaboration to real world in individuals. Elaboration is the cognitive process whereby individuals consciously or subconsciously establish paths between nodes in a semantic network representing newly learned material and nodes representing already known material [10]. In this sense, we will attempt to verify that

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 532–544, 2011.

CM is an effective methodology through an experiment in this paper. In this experimental, we examined the overall research question: "How do CM method and application domain knowledge influence performance on different types of schema understanding tasks?" So, we examine the effects of both CM and application domain knowledge on different types of schema understanding tasks: syntactic and semantic comprehension tasks and schema-based inferential problem-solving tasks. Our thesis was that while CM method knowledge is important in solving all such tasks, the role of application domain knowledge is contingent upon the type of understanding task under investigation. We use the theory of cognitive fit to establish theoretical differences in the role of application domain knowledge among the different types of schema understanding tasks.

2 Theory and Hypothesis

2.1 Semantic Network and Problem Solving

CM proposes that individuals will be more able to understand domain knowledge that complies with its criteria. To explain why its conditions affect analyst understanding, we must support its criteria with theories of cognition. Two bodies of theory help to explain this link. First, we draw upon semantic network theory to propose that CM lead analysts to construct efficient mental representations of a domain [11]. Semantic network theory states that individuals store concepts in memory as nodes connected by paths [12]. To perform cognitive tasks, individuals must recall concepts from memory. Recall follows a process of spreading activation: a node is primed in memory, which leads to paths connecting to it being activated [12]. Activation has to be strong enough for a search to reach a connected node. Empirical tests show that greater activation strength enables faster and more accurate recall [12]. CM leads to efficient mental representations by reducing activation strength and excluding relevant nodes.

Second, problem-solving theories suggest that the quality of a person's mental representation of a domain is a key driver of his/her ability to reason about the domain [13]. Specifically, problem solving theories suggest that a person reasons about a domain by drawing on his/her mental representation of the domain together with his/her mental representation of the problem s/he faces about the domain to construct a "problem space" in memory [13]. Tests show that problem solving performance is driven by a person's ability to search his/her problem space [13], [14]. Because semantic network theory suggests that CM lead to efficient mental representations, we can therefore tie CM to understanding by proposing that CM reduce analysts' ability to construct inefficient problem spaces in memory and thereby increase analysts' ability to search their problem space when reasoning about the domain.

An assumption of the preceding arguments is that individuals encode elements of CM into memory in a more or less one-to-one mapping so that causal relations in the CM will be manifest in an individuals' mental representation of the model. This assumption might not hold in two situations. The first is when individuals have plenty of time to analyze a model. In this case, individuals are likely to engage in elaboration processes to restructure their semantic network. Such elaboration not only alters the

mental representation of the model, but can improve an individual's memory by increasing the priming of concept nodes and improving the structure of the semantic network [12], [15]. The second exception is when individuals have existing knowledge of the domain represented in the model. In such cases, individuals internalize concepts in the model by integrating them with concepts in their existing semantic network of the domain [14], [16]. Thus, if individual describe concept nodes of casual relationships from CM, individuals may use their existing domain knowledge to identify problems in the model and infer or conclude a more plausible interpretation of the terms or relationships in the model when internalizing it [17]. Research on the influence of domain knowledge is complex and ongoing [18], but, complex and ongoing domain knowledge was described with CM.

2.2 Conceptual Schema Understanding Task

Our paper first present the types of conceptual schema understanding tasks that have been addressed in prior conceptual modeling research. Schema understanding tasks can be viewed as either read-to-do (with access to the schema) [18] or read-to-recall tasks (without access to the schema) [19]. Recall tasks have been used to investigate problem solvers' knowledge structures, that is, chunks of knowledge that are stored in internal memory and reused when appropriate. Perhaps the best known studies of this type have been those conducted by Weber (see, for example, [20]).

Two types of comprehension tasks that have been employed in prior IS research are supported in the education literature, which identifies two different types of knowledge, syntactic and semantic [21], [22]. We refer to such tasks as syntactic and semantic comprehension tasks. Syntactic knowledge involves understanding the vocabulary specific to a modeling formalism. Syntactic comprehension tasks are those that assess the understanding of just the syntax of the formalism associated with a schema. Semantic knowledge involves understanding the meaning, or the semantics, of the data embedded in the conceptual schema. Thus, semantic comprehension tasks are those that assess the understanding of the data semantics conveyed through constructs in the schema [23].

More recently, researchers have investigated tasks that require a deeper level of understanding than comprehension tasks, tasks that are referred to as problem-solving tasks [24]. We refer to a problem-solving task that can be solved using knowledge represented in the schema as a "schema based problem-solving task." The feature of using a CM is to support a "what-if" and "goal seeking" analysis. As revealed in previous studies, CMs must be further improved to deal with uncertainty and vagueness regarding the decision environments so that they may be used as a knowledge engineering tool to extract causal knowledge from factors representing environments [25]. For this purpose, a CM is organized as a matrix, in which it contains some specific inputs (or stimulus vectors) and produces outputs (or consequence vectors). The "what-if" and "goal seeking" analysis can be easily performed on this matrix representation. So, a further type of problem-solving task, which we refer to as an "inferential problem-solving task," requires CM users to use information beyond what is provided in the schema. A number of recent studies have used this type of task in addition to comprehension tasks [17], [18], [20], [26].

2.3 Cognitive Fit

The notion of task-technology cognitive fit is viewed as an important factor determining whether the use of technology would result in performance improvement [27], [28], [29], [30]. Briefly, the task-technology fit hypothesis argues that for an information system to have a positive impact on performance, it must be designed and utilized in such a way that it fits with the tasks it supports. When the information emphasized by the presentation matches the task, decision makers can use the same mental representation and decision processes for both the presentation and the task, resulting in faster and more accurate solutions [29]. When a mismatch occurs, one of two processes will occur. First, decision makers may transform the presented data to better match the task, which might increase the time needed and might decrease accuracy because any transformation can introduce errors [29]. Alternatively, decision makers may adjust their decision processes to match the presentation [31], decreasing accuracy and increasing time because the information does not match the ultimate needs of the task.

To better understand this relationship, we first need to explain the key concept equivocality of information. Daft and Macintosh defined information equivocality as "the multiplicity of meaning conveyed by information about organizational activities" (p. 211). High equivocality means confusion and lack of understanding [32]. Note that at times the literature uses the term equivocality to describe the characteristics of tasks. In this paper, we use the term exclusively to describe information characteristics. Moreover, we refer to less-analyzable task was consists of syntactic and semantic knowledge. By contrast, problem solving task was presented as analyzable task. Therefore, we will examine whether quality in decision making can be changed by the task type (analyzable vs less-analyzable) of Text and CM and its equivocality.

2.4 Hypotheses

Conceptual schemas are well-formalized representations of the structure of data within a specific application domain. However, CM knowledge has not been investigated in the context of conceptual schema understanding. Many researches on the role of CM knowledge have been conducted in the context of CM method development and application [6], [8], [9], [33]. But, many researches have been conducted in conceptual modeling area [15], [34], [35], [36]. [34] and [36] found that the quality of schemas developed by subjects with high-IS domain knowledge were generally superior to those developed by subjects with low-IS domain knowledge. Because CM was belong to IS domain knowledge, this paradigm was applied to CM knowledge. Therefore, we expect that decision makers with greater CM knowledge will perform better on all types of conceptual schema understanding tasks than those with less CM knowledge. We investigate the following hypotheses.

- H1 (a). Decision Makers with high CM knowledge are more accurate on syntactic comprehension tasks than those with low CM knowledge.
- H1 (b). Expert with domain knowledge are accurate on syntactic comprehension tasks than novice with domain knowledge.
- H2 (a). Decision Makers with high CM knowledge are more accurate on semantic comprehension tasks than those with low CM knowledge.

- H2 (b). Expert with domain knowledge are more accurate on semantic comprehension tasks than novice with domain knowledge.
- H3 (a). Decision Makers with high CM knowledge are more accurate on schema-based inferential problem-solving tasks than those with low CM knowledge.
- H3 (b). Expert with domain knowledge are more accurate on inferential problem-solving tasks than novice with domain knowledge.

3 Overview of the Experiments

To test the hypotheses, we conducted an experiment. The design of the experiment was motivated in part by research undertaken by Mayer and his colleagues [16], [37]. They had performed a series of experiments in which they had examined the impact of text versus diagrams on their participants' ability to learn a domain. Mayer and his colleagues made three predictions. First, they hypothesized that participants who received diagrams would perform better on problem-solving tasks than those who received text. They argued that diagrams allowed participants to develop more sophisticated cognitive models of the domain to be learned. Participants would develop a deep understanding of the domain rather than a surface understanding. Second, they hypothesized that participants who received diagrams would perform worse on verbatim-recall tasks than those who received text. They argued that the deeper conceptual processing undertaken by participants who received the diagrams would undermine their ability to retain the information they needed for verbatim recall. Third, they hypothesized that participants who received diagrams would perform about the same on comprehension tests as those who received text. To the extent that comprehension tests required more conceptual understanding of the domain, participants who received diagrams would do better.

This experiment involved a comprehension task and problem solving task. The comprehension questions we asked of participants in the experiment were straightforward and often relied on analysis of features of the causal map. The problem solving task was that can be solved using knowledge represented in the schema as a "schema based problem-solving task." A further type of problem-solving task, which we refer to as an "inferential problem-solving task," requires conceptual modelers to use information beyond what is provided in the schema [38]. Thus, our expectation was that participants who received the familiarity with the application domain knowledge would still outperform those who received the unfamiliarity with the application domain knowledge.

4 Experiment

We conducted a laboratory experiment to test the hypotheses present above. The experiment employs two-way ANOVA of 2 x 2 design. First, the between-subject factor is the type of participant group: expertise (Master Black Belt: MBB, Black Belt: BB) or novice (Green Belt: GB). The expert group comprised of MBBs, BBs who had received fairly extensive training to undertake 6-sigma analysis and evaluation during

more than 3years; the novice group comprised of GB who had an experience of 6-sigma analysis and evaluation for 1~6 months. Second, the between-subject factor is the CM method knowledge type (high CM knowledge vs. low CM knowledge).

4.1 Task Material

A real world 6-sigma project analysis and evaluation case was used as the experimental stimulus. In order to provide the subjects with a realistic environment, the CM case was selected based on three criteria. First, the CM case must describe fundamental problems in the 6-sigma analysis and evaluation project that can only be used by MBBs, BBs and GBs. Second, the case should balance the amount of relevant information among multiple nodes and causality from different perspectives. Third, the scope of the case should not be exceedingly large, since the subjects must be able to understand the entire 6-sigma business system within the given time limits of the experiment (Appendix A).

To build a causal map of 6-sigma CM, we used a four step process suggested by [33]. To determine the factors regarding 6-sigma project analysis and evaluation, an extensive open interview was performed with 5 experts such as MBBs. Our interview used open interview techniques with probes to facilitate the interview process [1]. Respondents were asked about 6-sigma project analysis and evaluation factors that seem relevant to the 6-sigma project analysis and evaluation decision process. The interviewers did not constrain responses to questions. Each interview lasted from 60 to 90 minutes.

To form the groups, we used participants' scores on the CM method test that the participants. For this, we divided the participants with two groups; expert and novice. One group was educated for CM method during two or three days and had to answer several questions by using CM. The other group was simply educated for CM method only to do an experiment during half an hour without special education.

We applied same dependent variables, syntactic and semantic comprehensive as well as schema of inferential problem-solving task.

4.2 Participants

The participants were MBBs, BBs who work on 6-sigma project during more than 3 years, and GBs work on 6-sigma project for 1-6 months. The 64 responses, totally, were used to prove research hypotheses. To investigate our hypotheses related to CM method knowledge, we needed to form groups of participants with high and low expertise in the CM method. Participation was voluntary; all participants were offered $30 gift certificates to encourage their participation in the experiment. All the respondents were male. On average, expert group was 41.3 years old, and they conducted the 6-sigma related project 6.4 unit per year, but novice group was 0.8 unit per year.

4.3 Experimental Procedures

The experimental procedures were divided into three sections. First, the subjects were given instructions about the general nature of the experiment and were told that verbal

protocols would be collected. Second, one of the researchers provided a review of causal map method concepts to the subjects for approximately 20~30 minutes. Third, the subjects were then presented with the experimental 6-sigma CM and were asked to analysis and evaluation the project based on the 6-sigma CM. Since the 6-sigma CM was provided on paper, subjects could spread them out physically in parallel. The subjects were given 60 minutes to understand the diagrams and come up with what they believed were the fundamental problems in the 6-sigma project analysis and evaluation. Due to the intended complexity of the experimental task, most subjects worked up until the 60- minute time limit.

4.4 Results

The sample size was 64 for the analysis on understanding tasks questionnaire. Data associated with understanding tasks was analyzed using a MANOVA test with the three independent variables, expert-novice group and High-Low CM knowledge type. The MANOVA model was applied to test the influence of the Expert-Novice type (Expert-Group vs Novice-Group) and CM knowledge type (High CM Knowledge-Group vs Low CM Knowledge-Group) on the three dependent variables. The main effect of the Expert-Novice type was significant ($F(3, 58) = 34.059$, $p=0.000$) and the main effect of the CM knowledge type was also significant ($F(3, 58) = 7.170$, $p=0.000$). The interaction effect between the Expert-Novice type and CM knowledge type was significant ($F(3, 58) = 3.279$, $p=0.027$, $p<0.05$) was significant. Because the MANOVA results were significant, these results were further analyzed using individual ANOVAs to examine the effects of the Expert-Novice type and CM knowledge type on each dependent variable.

Table 1. Results of the ANOVA for Schema Understanding Tasks

Syntactic Comprehension	DF	Mean Squared	F-value	p
Expert-Novice Type	1	0.563	49.724	0.000***
CM Knowledge Type	1	0.181	15.967	0.000***
CM Knowledge Type x Expert-Novice	1	0.016	1.381	0.245
Error	60	0.011	-	-
Semantic Comprehension	**DF**	**Mean Squared**	**F-value**	**P**
Expert-Novice Type	1	0.526	93.100	0.000***
CM Knowledge Type	1	0.051	8.967	0.004**
CM Knowledge Type x Expert-Novice	1	0.022	3.985	0.050*
Error	60	0.006	-	-
Inferential Problem-Solving	**DF**	**Mean Squared**	**F-value**	**P**
Expert-Novice Type	1	0.581	21.263	0.000***
CM Knowledge Type	1	0.263	9.606	0.003**
CM Knowledge Type x Expert-Novice	1	0.114	4.166	0.046*
Error	60	0.027	-	-

Table 1 reports the results of ANOVA. In syntactic comprehension, the main effect of the Expert-Novice type ($F(1, 60) = 49.724$, $p=0.000$) and CM knowledge type ($F(1, 60) = 15.967$, $p=0.000$) was significant. But, the interaction effect between the Expert-Novice type and CM knowledge type was not significant ($F(1, 60) = 1.381$, $p=0.245$). In semantic comprehension, the main effect of the Expert-Novice type ($F(1, 60) = 93.100$, $p=0.000$) and CM knowledge type ($F(1, 60) = 8.967$, $p=0.004$) was significant. The interaction effect between the Expert-Novice type and CM knowledge type was also significant ($F(1, 60) = 3.985$, $p=0.050$). In inferential problem solving, the main effect of the Expert-Novice type ($F(1, 60) = 21.263$, $p=0.000$) and CM knowledge type ($F(1, 60) = 9.606$, $p=0.003$) was significant. The interaction effect between the Expert-Novice type and CM knowledge type was also significant ($F(1, 60) = 4.166$, $p=0.046$).

The mean values and standard deviations are shown in Table 2, for each type of understanding task for participants with high and low-CM knowledge in each of the expert and novice group, respectively.

Table 2. T-Test Result for Performance on Schema Understanding Tasks

(a) CM Knowledge Type

Understanding Task Type	CM Knowledge Type		t-Value for difference (Hi-CM vs Lo-CM)	p-value
	Hi-CM (16)	Lo-CM (16)		
Syntactic comprehension	0.87 (0.15)	0.76 (0.14)	2.98	0.004**
Semantic comprehension	0.87 (0.13)	0.82 (0.11)	2.08	0.05*
Inferential problem solving	0.91 (0.10)	0.78 (0.25)	2.64	0.02*

(b) Expert-Novice Type

Understanding Task Type	Expert-Novice Type		t-Value for difference (Expert vs Novice)	p-value
	Expert Group (16)	Novice Group (16)		
Syntactic comprehension	0.91 (0.09)	0.73 (0.14)	6.31	0.000***
Semantic comprehension	0.93 (0.07)	0.75 (0.10)	8.90	0.000***
Inferential problem solving	0.94 (0.06)	0.75 (0.24)	4.23	0.000***

In Table 2, participants with High-CM-knowledge group performed better than those with Low-CM-knowledge group for all types of understanding tasks; syntactic comprehension, semantic comprehension, and inferential problem-solving tasks, in both Expert group and Novice group. For H1(a), decision makers with high CM knowledge are more accuracy than those with low CM knowledge (0.87 for Hi-CM knowledge and 0.76 for Low CM knowledge; $t = 2.99$, $p=0.004$), in both Expert-Novice groups. For H2(a), decision makers with high CM knowledge are more accuracy than those with low CM knowledge (0.87 for Hi-CM knowledge and 0.81 for Low CM knowledge; $t = 2.08$, $p=0.050$), in both Expert-Novice groups. For H3(a), decision makers with high CM knowledge are more accuracy than those with low CM

knowledge (0.91 for Hi-CM knowledge and 0.78 for Low-CM knowledge; t = 2.64, p=0.020), in both Expert-Novice group. Therefore, all differences were significant influence of CM knowledge on the three comprehension task in the both Expert-Novice groups. Hence, overall, results in both in both Expert-Novice groups suggest that CM knowledge affects problem-solving performance on all types of understanding tasks, supporting Hypotheses 1(a), 2(a), and 3(a).

Next, we investigate Hypotheses 1(b), 2(b), and 3(b), that the influence of Expert and Novice type on performance in understanding tasks is contingent on the type of understanding task under investigation. For H1(b), expert are more accuracy than novice (0.91 for Expert and 0.73 for Novice; t = 6.31, p=0.000). For H2(b), expert are more accuracy than novice (0.93 for Expert and 0.75 for Novice; t = 8.90, p=0.000). For H3(b), expert are more accuracy than novice (0.94 for Expert and 0.75 for Novice; t = 4.23, p=0.000). Therefore, all differences were significant influence of Expert-Novice on the three comprehension task in the both CM knowledge types. Hence, overall, results suggest that Expert-Novice affects problem-solving performance on all types of understanding tasks, supporting Hypotheses 1(b), 2(b), and 3(b).

5 Discussion and Concluding Remarks

In this section we discuss our findings and present the contributions of our research. We conclude with the implications of our research for both future research and for practice.

In the experiment, we explored the role of CM knowledge and application domains in decision making with CM method. We examined the overall research question: "How do CM method and application domain knowledge influence performance on different types of schema understanding tasks?" To do so, we conducted an experiment in which we manipulated both CM knowledge and application domain knowledge. This research addresses the role of CM knowledge and application domain knowledge in understanding decision problem both theoretically via the theory of cognitive fit, and empirically. Specifically, we address the role of CM knowledge and application domain knowledge on the performance of decision makers on different types of understanding tasks. Our research shows that while CM knowledge is important to the solution of all types of schema understanding tasks, application domain knowledge affects the solution of just inferential problem solving tasks, tasks for which decision makers must transform knowledge with CM method in the schema into a form suitable for task solution.

The experiment has the following limitations. First, our study was conducted in a laboratory setting, which means that it suffered the typical limitations of all experiments. On the other hand, we were able to control for many aspects that might have come into play had we conducted our study in a professional setting. Second, we presented the tasks to participants in the same order (syntactic, semantic, inferential problem solving) so that we could have greater confidence that any effects of application domain knowledge on the more demanding inferential problem-solving tasks were due to application knowledge itself, and not to lack of knowledge of the schema. Because we presented our tasks always in the same sequence, we need to consider the potential effects of fatigue. Note that in the presence of fatigue effects, performance in

the second domain would have been worse than in the first domain for each of the inferential problem-solving tasks, a phenomenon we did not observe. Further, subjects were given different types of tasks in the same order in each application domain; thus, in comparing performance across CM knowledge and applications domains, the effects of fatigue were essentially controlled.

Acknowledgment. This research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

[1] Axelrod, R.: Structure of Decision: The Cognitive Maps of Political Elites. Princeton University Press, Princeton (1976)
[2] Hart, J.A.: Comparative cognition: politics of international control of the oceans. In: Axelrod, R. (ed.) Structure of Decision. Princeton University Press, Princeton (1976)
[3] Hart, J.A.: Cognitive maps of three Latin American policy makers. World Politics 30(1), 115–140 (1977)
[4] Robert, F.S.: Strategy for the energy crisis: the case of commuter transportation policy. In: Axelrod, R. (ed.) Structure of Decision: The cognitive Maps of Political Elites. Princeton University Press, Princeton (1976)
[5] Montazemi, A.R., Conrath, D.W.: The Use of Cognitive Mapping for Information Requirements Analysis. MIS Quarterly 10(1), 45–56 (1986)
[6] Lee, K.C., Kwon, S.J.: The use of cognitive maps and case-based reasoning for B2B negotiation. Journal of Management Information Systems 22(4), 337–376 (2006)
[7] Clarke, L., Mackaness, W.: Management 'Intuition': An Interpretative Account of Structure and Content of Decision Schemas Using Cognitive Maps. Journal of Management Studies 38(2), 147–172 (2001)
[8] Satur, R., Liu, Z.Q.: A Contextual Fuzzy Cognitive Map Framework for Geographic Information Systems. IEEE Transactions on Fuzzy Systems 7(5), 481–494 (1999)
[9] Liu, Z.Q., Satur, R.: Contextual fuzzy cognitive map for decision support in geographic information systems. IEEE Transactions on Fuzzy Systems 7(5), 495–507 (1999)
[10] Bradshaw, G.L., Anderson, J.R.: Elaborative encoding as an explanation of levels of processing. J. Verbal Learning and Verbal Behavior 21, 165–174 (1982)
[11] Collins, A.M., Quillan, M.R.: Retrieval time from semantic memory. J. Verbal Learn. Behavior 8, 240–247 (1969)
[12] Ashcraft, M.H.: Cognition. Prentice Hall, Upper Saddle River (2002)
[13] Newell, A., Simon, H.A.: Human Problem Solving. Prentice Hall, Englewood Cliffs (1972)
[14] Pretz, J.E., Naples, A.J., Sternberg, R.J.: Recognizing, defining, and representing problems. In: Davidson, J.E., Sternberg, R.J. (eds.) The Psychology of Problem Solving, pp. 3–30. Cambridge University Press, Cambridge (2003)
[15] Weber, R.: Are attributes entities? A study of database designers' memory structures. Inform. Systems Res. 7, 137–162 (1996)
[16] Mayer, R.: Models for understanding. Rev. Educational Res. 59, 43–64 (1989)
[17] Burton-Jones, A., Weber, R.: Understanding relationships with attributes in entity-relationship diagrams. In: Proc. 20th International Conference Information Systems, Charlotte, NC, pp. 214–228 (1999)

[18] Khatri, V., Ramesh, V., Vessey, I., Clay, P., Park, S.J.: Understanding conceptual schemas: Exploring the role of application and IS domain knowledge. Inform. Systems Res. 17(1), 81–99 (2006)
[19] Burkhardt, J.M., Détienne, F., Wiedenbeck, S.: Object-oriented program comprehension: Effect of expertise, task and phase. Empirical Software Engineering 7(2), 115–156 (2002)
[20] Bodart, F., Sim, M., Patel, A., Weber, R.: Should optional properties be used in conceptual modelling? A theory and three empirical tests. Information Systems Research 12(4), 385–405 (2001)
[21] Schneiderman, B., Mayer, R.E.: Syntactic/semantic interactions in programmer behavior: A model and experimental results. Internat. J. Comput. Inform. Sci. 8, 219–238 (1979)
[22] Mayer, R.E.: Thinking, Problem Solving, Cognition, pp. 560–578. W.H. Freeman and Company, New York (1991)
[23] Elmasri, R., Navathe, S.B.: Fundamentals of Database Systems, 2nd edn. Benjamin/Cummings Publishing Co., Redwood City (1994)
[24] Gemino, A.: Empirical comparisons of animation and narration in requirements validation. Requirements Engineering 9(3), 153–168 (2004)
[25] Taber, R.: Knowledge processing with fuzzy cognitive maps. Expert Systems with Applications 2(1), 83–87 (1991)
[26] Burton-Jones, A., Meso, P.N.: Conceptualizing Systems for Understanding: An Empirical Test of Decomposition Principles in Object-Oriented Analysis. Information Systems Research 17(1), 38–60 (2006)
[27] Goodhue, D.L., Thompson, R.L.: Task Technology Fit and Individual Performance. MIS Quarterly 19, 213–236 (1995)
[28] Tan, J.K.H., Benbasat, I.: The Effectiveness of Graphical Presentation for Information. Decision Sciences 24, 167–191 (1993)
[29] Vessey, I.: Cognitive Fit: A Theory-Based Analysis of the Graphs Versus. Decision Sciences 22, 219–240 (1991)
[30] Vessey, I., Galletta, D.F.: Cognitive Fit: An Empirical Study of Information Acquisition. Information Systems Research 2, 63–84 (1991)
[31] Perrig, W., Kintsch, W.: Propositional and situational representations of text. Journal of Memory and Language 24, 503–518 (1985)
[32] Daft, R.L., Lengel, R.H., Trevino, L.K.: Message Equivocality, Media Selection and Manager Performance. MIS Quarterly (11), 355–364 (1987)
[33] Nelson, K.M., Nadkarni, S., Narayanan, V.K., Ghods, M.: Understanding software operations support expertise: a revealed causal mapping approach. MIS Quarterly 24(3), 475–507 (2000)
[34] Batra, D., Davis, J.G.: Conceptual data modelling in database design: Similarities and differences between expert and novice designers. Internat. J. Man-Machine Stud. 37, 83–101 (1992)
[35] Lee, H., Choi, B.G.: A comparative study of conceptual data modeling techniques. J. Database Management 9, 26–35 (1998)
[36] Moody, D.L., Shanks, G.G., Darke, P.: Improving the quality of entity relationship models—Experience in research and practice. In: Ling, T.-W., Ram, S., Li Lee, M. (eds.) ER 1998. LNCS, vol. 1507, pp. 255–276. Springer, Heidelberg (1998)
[37] Mayer, R.E., Gallini, J.K.: When is an illustration worth a thousand words. J. Ed. Psych. 82, 715–726 (1990)
[38] Gemino, A., Wand, Y.: Evaluating modeling techniques based on models of learning. Communications of ACM 46(10), 79–84 (2003)

Appendix A

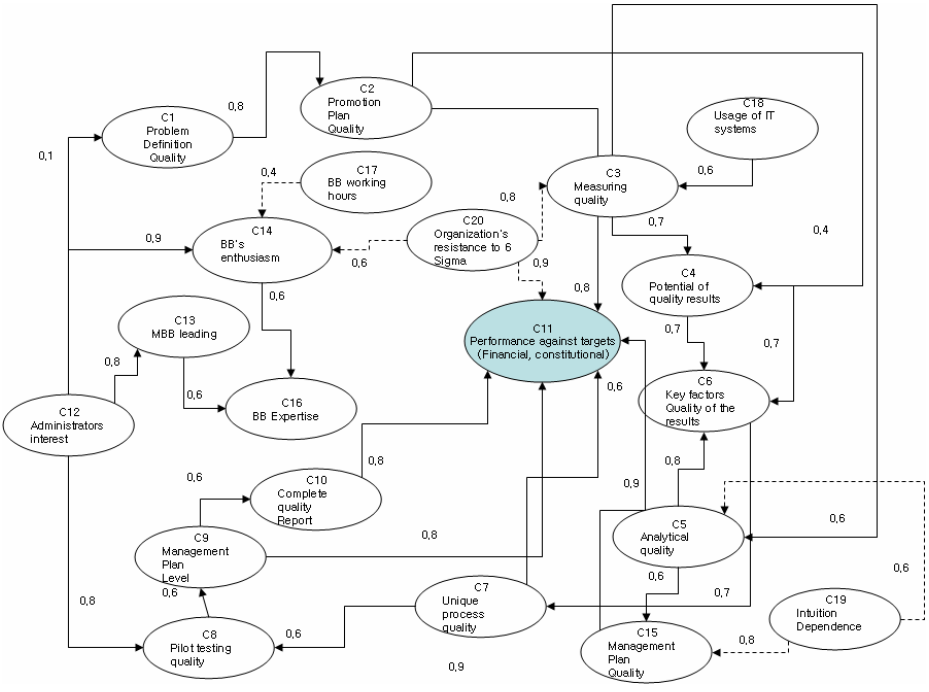
Appendix A 6-Sigma project CM

The figure below shows the result derived from interview with professionals on 6-sigma project development and evaluation of performance. The result consists of factors that were configured into the map. Appendix_Table 1 shows the values for the input node on the map. Appendix_Figure 1 is a map that proposing various reasoning once a user enters a value.

Appendix_Table 1. Development and evaluation of the input node values in the map for 6-sigma project

Organization's resistance to 6-sigma		Usage of IT System	
No conflict	0.4	High	1.0
A bit of conflict	0.6	Average	0.8
A lot of conflict	0.8	Low	0.6
Labor disputes	1.0	-	-

Organizational Interest		Working hours	
Low	0.2	1 hour commitment	0.5
Average	0.5	2 hours commitment	0.6
High	0.8	3 hours commitment	0.8
Very High	1.0	4 hours commitment	1.0



inference	1 Problem Definition Quality	2 Promotion Plan Quality	3 Measuring quality	4 Potential of Quality results	5 Analytical quality	6 Key factors Quality of the results	7 Unique Process Quality	8 Pilot testing quality	9 Management Plan Level	10 Complete quality Report	11 Performance against targets (Financial, constitution	12 Administrators interest	13 MBB leading	14 BB's enthusiasm	15 Management Plan Quality	16 BB Expertise	17 BB working hours	18 Usage of IT Systems	19 Intuition Dependence	20 The organization's resistance to 6 Sigma
Initial value(C0)	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C3	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
C3 * AM	0,1	0,1	0,0	0,0	1,2	2,4	3,3	3,4	2,6	2,1	10,1	0,0	0,8	0,9	1,3	2,2	0,0	0,0	-1,9	0,0
1/2 threshold	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0
C4	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
C4 * AM	0,1	0,1	0,0	0,0	1,2	2,4	3,3	3,4	2,6	2,1	10,1	0,0	0,8	0,9	1,3	2,2	0,0	0,0	-1,9	0,0
1/2 threshold	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0

Appendix_Fig. 1. Analysis and evaluation of Cognitive Map in 6-Sigma Project

Antecedents of Team Creativity and the Mediating Effect of Knowledge Sharing: Bayesian Network Approach to PLS Modeling as an Ancillary Role

Kun Chang Lee[1], Dae Sung Lee[2,*], Young Wook Seo[3], and Nam Young Jo[4]

[1] Professor of MIS and WCU Professor of Creativity Science, SKK Business School and Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2] PhD Candidate, SKK Business School, Sungkyunkwan University, Seoul 110-745, Republic of Korea
Tel.: +82 2 760 0505; Fax: +82 2 760 0440
leeds1122@gmail.com, leeds@skku.edu
[3] Researcher, Software Engineering Center at NIPA, Seoul 138-711, Republic of Korea
seoyy123@gmail.com
[4] PhD Candidate, SKK Business School, Sungkyunkwan University, Seoul 110-745, Republic of Korea
higlobe@naver.com

Abstract. Central points of this paper are placed on team creativity and Bayesian network approach as an assistant means. Above all, we seek to intensify our PLS model by using Bayesian network (BN) approach as an ancillary role. Beyond managers' control, we emphasize a voluntarily and informally emergent structure and introduce a social network perspective within team creativity. In this sense, we propose a new integrative team creativity model in which shared leadership, interpersonal trust and knowledge sharing are included and their subsequent influence on team creativity is analyzed. For the sake of empirical analysis, an e-learning course was administered in a private university, and 40 teams were organized for this study. 249 valid questionnaires were garnered, and initially analyzed by PLS (Partial least squares) model. Then, we suggested a new PLS model based on the results of Bayesian networks, and confirmed the successful application of our proposed approach.

Keywords: Team creativity, Bayesian networks, Shared leadership, Social network, Interpersonal trust, Knowledge-sharing, PLS (Partial least squares) model.

1 Introduction

The ability to develop and implement innovation is essential to today's competitive business environment [19]. Many organizations have turned to team-based work

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 545–555, 2011.

systems to increase their responsiveness and ability to foster innovation (cf. [27]). Such organizations need to be concerned not only with fostering creativity and innovation among individual employees, but also developing creative and innovative teams. However, fewer studies have investigated "team creativity" compared to those studying individual creativity. In this article, we explore the effects of shared leadership, knowledge-sharing, and interpersonal trust on team creativity taking a team-based approach. Broadly stated, Bayesian networks (BNs) are directed acyclic graphs (DAGs) with a set of probability tables. According to Lauritzen [21], graphical models, in particular those based on DAGs, have natural causal interpretations and thus form a language in which causal concepts can be discussed and analyzed in precise terms. Therefore, BNs are extremely valuable for providing actionable information and advice. This article proposes to implement Bayesian networks (BNs) after conducting an initial PLS modeling. Bayesian network approach, as an ancillary role, will help us to find out unexpected results and revise the initial PLS model. Therefore, an empirical study for our integrative team creativity model is presented to demonstrate the successful application of the proposed method.

2 Initial PLS Model

2.1 Theoretical Background and Hypotheses

The research model used in this paper builds on the team level constructs of shared leadership, cognition-based trust, affect-based trust, knowledge sharing, and their subsequent impacts on team creativity.

Shared leadership and Team creativity

Intrinsic motivation is a key factor in creativity [2] and is often considered the mechanism by which situational factors such as leadership contribute to creativity [2, 30]. Transformational leadership is positively related to follower creativity [35]. However, prior research has highlighted the analysis of vertical leadership and individual creativity. Our model requires another kind of leadership as an emergent team property to analyze team creativity. Shared leadership is a relational phenomenon that involves mutual influence among team members [26]. Social network theory provides a natural theoretical and analytical approach to studying the relational influence structures of teams [26]. Therefore, although there are a few useful self-reported ratings [5, 32] that can measure shared leadership, this article focuses on a social network approach [24, 7] that uses density, a measure of the total amount of leadership displayed by team members as perceived by others on the team. Equation (1) shows how to calculate density for shared leadership. On the basis of these studies, we proposed the following hypothesis:

H1: Shared leadership will positively contribute to team creativity.

$$\text{Density} = S / 7N(N-1) \quad (1)^1$$

[1] In this equation, S is the sum of all values that team members would rate each other for leadership. N equals the number of team members; N (N-1) is the total number of possible ties in a team. The number 7 represents the maximum value rated by a peer in a team.

Shared leadership and Trust

Podsakoff et al. [33] showed that trust, conceptualized as faith in and loyalty to a leader is directly related to transformational leadership. However, prior research has focused on a single leader and has excluded reciprocal trust. Social network analysis is suitable for the study of multiple sources of leadership influence and the distributed leadership grounded in the interactions among team members [26, 13]. Moreover, interpersonal trust (cognition- and affect-based trust) is appropriate for the social network approach [12]. On the basis of these studies, we proposed the following hypotheses:

H2: Shared leadership will positively contribute to cognition-based trust.
H3: Shared leadership will positively contribute to affect-based trust.

Shared leadership and Knowledge sharing

Knowledge sharing is a critical team process because if knowledge is not shared, the cognitive resources available to a team remain underutilized [4]. The team leader has an important role to play in properly extracting the shared knowledge of the team. Empowering leadership is different from autocratic leadership, and one of the central differences in the outcomes is that autocratic leadership inhibits knowledge sharing among team members [39]. Studies have shown that team communication styles, agreeable and extravert, are positively related to knowledge sharing willingness and behaviors [15]. Therefore, leadership distributed within a team, or shared leadership, may be associated with knowledge sharing. From these findings, we proposed the following hypothesis:

H4: Shared leadership will positively contribute to knowledge sharing.

Trust and Knowledge sharing

Researchers have used social exchange theory to examine how trust and justice, two key components in interpersonal relationships, relate to knowledge sharing [31, 34]. Examining trust and justice is important because knowledge sharing involves providing knowledge to another person or a collective unit, such as a team or community of practice with expectations for reciprocity (e.g., [38]). Research has shown that affect- and cognition-based trust have a positive influence on knowledge sharing at the dyadic and team levels [11, 28, 38]. On the basis of these studies, we proposed the following hypotheses:

H5: Cognition-based trust will positively contribute to knowledge-sharing.
H6: Affect-based trust will positively contribute to knowledge-sharing.

Knowledge sharing and Team Creativity

Employees may share their ideas with others to further develop them and to facilitate creativity [29]. An effective working relationship exists when both parties exchange knowledge resources to foster progress and to resolve difficulties of both technical and artistic natures. Sharing knowledge is an important facilitator of creative collaboration [23]. These findings suggest the following hypothesis:

H7: Knowledge sharing will positively contribute to team creativity.

Trust and Team Creativity

Even though a general consensus holds that mutual trust is a key factor in joint innovative developments, the literature on the effects of trust on joint/team creativity remains largely inconclusive [6]. On the other hand, several authors [17, 36] have observed that mutual trust is conducive to an increase in joint/team creativity. These findings suggest the following hypothesis:

H8: Cognition-based trust will positively contribute to team creativity.
H9: Affect-based trust will positively contribute to team creativity.

2.2 Empirical Analysis

Empirical data were gathered from a questionnaire survey. The participants were undergraduate students who had taken a web-based e-learning course called "Digital information technology and its application" at a Korean University. We created 40 teams consisting of four to eight members in an impersonal situation and gave each

Table 1. The Results of Confirmatory Factor Analysis for Initial PLS Model

Measure	AVE	Composite Reliability	R Square	Cronbach's Alpha
Shared Leadership ‡	N/A	N/A	N/A	N/A
Cognition-based Trust	0.912	0.969	0.590	0.952
Affect-based Trust	0.862	0.962	0.509	0.947
Knowledge-Sharing	0.825	0.950	0.695	0.930
Team Creativity	0.864	0.950	0.807	0.921

‡ A measure of the total amount of leadership displayed by team members as perceived by others on a team.

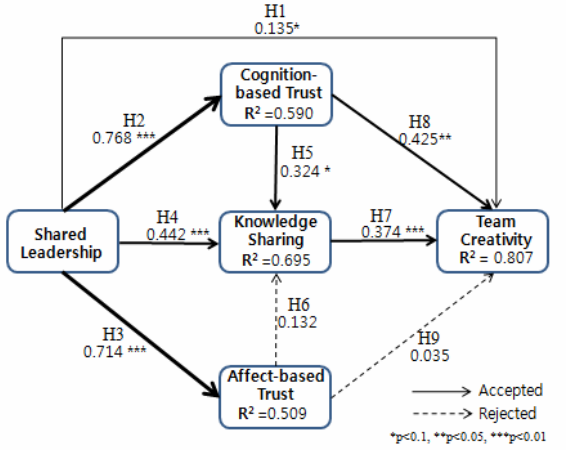


Fig. 1. Initial PLS Model

team an assignment that required creativity, notifying the participants that the assignment would have a critical effect on their grade. Finally, we provided questionnaires to students in the course, and yielded 249 useable cases. Then, we used SmartPLS 2.0 to analyze the measurement and structural models. We undertook assessments of discriminant and convergent validities, and secured them (Table 1). As shown in Fig. 1, our model results show that H6 and H9 are rejected and the other hypotheses are accepted.

3 Revised PLS Model

3.1 Bayesian Networks

Bayesian Networks

Bayesian networks (BNs) are graphical models that combine elements of both graph and probability theory. Broadly stated, BNs are directed acyclic graphs (DAGs) with a set of probability tables. A BN encodes the probability distribution of a set of random variables by specifying a set of conditional independence assumptions together with a set of relationships among these variables and their related joint probabilities [20]. In this paper, we used WEKA which offers various algorithms such as: hill climbing, K2, simulated annealing, genetic, tabu, TAN, and so on. We selected the TAN (Tree Augmented Naïve Bayes) among these algorithms. It can produce a causal-effect graph in which the class attribute treated as the only and greatest parent node of all other nodes is located at the top in the DAG [16]. The causal-effect graph of the TAN is formed by calculating the maximum weight spanning tree using Chow and Liu's method [10]. The TAN is an extension of the Naïve Bayes: it removes the Naïve Bayes assumption that all the attributes are independent. Moreover, the TAN finds correlations among the attributes and connects them in the network structure learning process [37]. According to Friedman et al. [16], the TAN provides for additional edges between attributes that capture correlations among them, and it approximates the interactions between attributes by using a tree structure imposed on the Naïve Bayes structure. Bayesian network classifiers incorporated in WEKA, such as the Bayesian network with the TAN search algorithm, have exhibited excellent performance in data mining [8].

BNs Experiment

The Bayesian network classifier with the TAN search algorithm was implemented with WEKA. We selected the "equal frequency" option under unsupervised discretization in WEKA. In our dataset, most items of latent variables have skewed distributions for the positive direction. Therefore, in case of equal width discretization, lower values of the items have very few numbers that make discretization effects to be meaningless. The prediction accuracy of this experiment was 70%. Consequently, casual relationships are able to be easily identified using the TAN search algorithm. Compared to the initial PLS model, the results of Bayesian network show unexpected relationships (Table 2). As shown in Fig. 2, when the team creativity value is set to the highest level, the posterior probability for each high level of all the variables increased. On the other hand, the medium level of affect-based trust value is the most

outstanding. This indirectly explains the rejected hypothesis 9 the initial PLS model, in that the causal relationship doesn't represent the linearity between affect-based trust and team creativity. As mention above, the previous researches [14, 9] did not find a direct and positive impact of mutual trust on the creativities of R&D project teams.

Table 2. Comparison BN with Initial PLS Model

Path	Causality relationship in BN		Initial PLS Model	
			Hypotheses	results
Cognition-based Trust → Affect-based Trust	Yes	New	N/A	N/A
Knowledge Sharing → Cognition-based Trust		Reverse to PLS	H5	Accepted
Shared Leadership → Affect-based Trust	No		H3	Accepted
Shared Leadership → Knowledge Sharing			H4	Accepted
Affect-based Trust → Knowledge Sharing			H6	Rejected

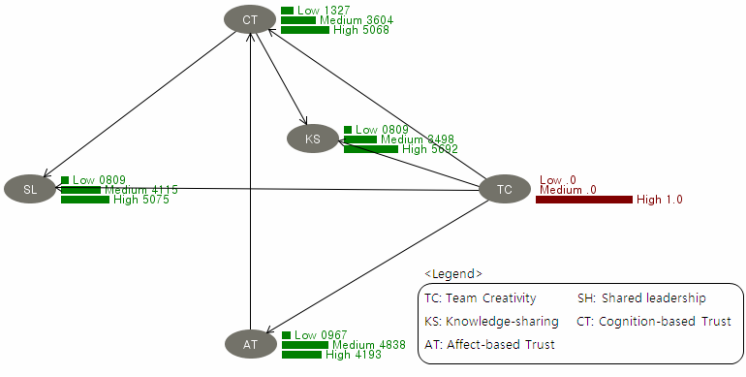


Fig. 2. Bayesian Network Analysis

3.2 Proposed Approach

Comparison Bayesian Networks with PLS modeling

Anderson and Vastag [3] stress that the SEM (Structural Equation Model) such as PLS model is likely the preferred method if the objective is only a description of theoretical constructs with no interest in inference to observable variables, while the Bayesian network approach should be selected if objectives include prediction and diagnostics of observed variables. The SEM highlights theory confirmation while Bayesian network approach stresses causal explanation [20]. Similar to PLS path modeling, BNs graphically portray the nature and strength of relationships—not necessarily hypothesized—among several constructs or variables. In a BN, a directed acyclic graph (DAG) represents a set of conditional independence constraints, or assumptions, among a given number of variables and their related conditional probability distributions. Instead of formulating a network, or model, on the basis of theory and then testing it, as with the PLS path modeling, BN techniques identify the network that fits the data best.

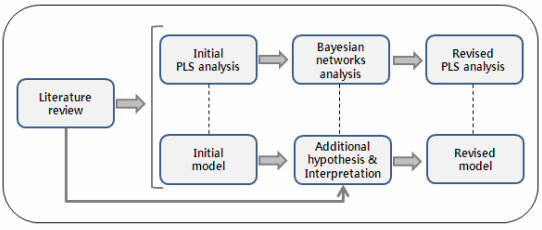


Fig. 3. Proposed approach for causal analysis

Application of Proposed Approach

Regarding theory dependency, the Bayesian network is data driven with no restrictions, while PLS path modeling is based on theory [20]. A few studies have suggested implementing the Bayesian network prior to conducting PLS path modeling for the solution of excessively complicated relationships between constructs [37]. However, without previous theories, serious problems would arise in social science research. As shown in Fig. 3, we propose a new method to revise PLS modeling according to the DAG of the Bayesian network. The proposed approach for causal analysis consists of four phases: 'Literature review', 'Initial PLS analysis, 'Bayesian network analysis' and 'Revised PLS analysis'. We employ WEKA to obtain a DAG through Bayesian network classifiers with a TAN search algorithm. Based on the DAG, we amended the initial model and implemented analysis with SmartPLS.

Additional Hypothesis and Interpretation for Revised PLS model

Table 2 and Fig. 2 show that there is an unexpected path ('Path CT → AT') that was not considered in the initial model. We should additionally consider this path in PLS model. Cognitive trust (CT) provides a base for affective trust (AT) and should therefore exist before affective trust develops [22]. But as affective trust matures, the potential for decoupling of trust dimensions and reverse causation increases [25]. Attitude theory researchers have long argued that the relationship between cognition and affect in attitude formation is bidirectional [18]. On the basis of these studies, we proposed the following hypotheses and amended the initial model (Fig. 4):

H10: Cognition-based trust will positively contribute to affect-based trust.

There are three paths that have no causal relationship in the DAG of the Bayesian network (Fig. 2). Two of them ('Path SL → AT' and 'Path AT → KS') are consistent with the results of the revised PLS model (Fig. 4). On the other hand, the 'Path SL → KS' is inconsistent with the results of the revised PLS model. Non-causal relationship of the Path (H4) is due to the nonlinearity of Bayesian networks. In other words, Shared leadership doesn't directly contribute to knowledge sharing, but indirectly through cognition-based trust in Bayesian network. The 'Path CT →KS' in the initial PLS model has the reverse direction (KS → CT) in the DAG of the Bayesian network. Trust is an attitude construct, while knowledge sharing is a behavioral construct. We cannot find any previous study employing knowledge sharing as an antecedent on trust. TRA (Theory of reasoned action) posits that individual behavior is driven by behavioral intentions where behavioral intentions are a function of an individual's attitude toward the behavior and subjective norms surrounding the

performance of the behavior [1]. Attitude toward the behavior is defined as the individual's positive or negative feelings about performing a behavior. Therefore, it is difficult to use knowledge sharing as an antecedent of trust.

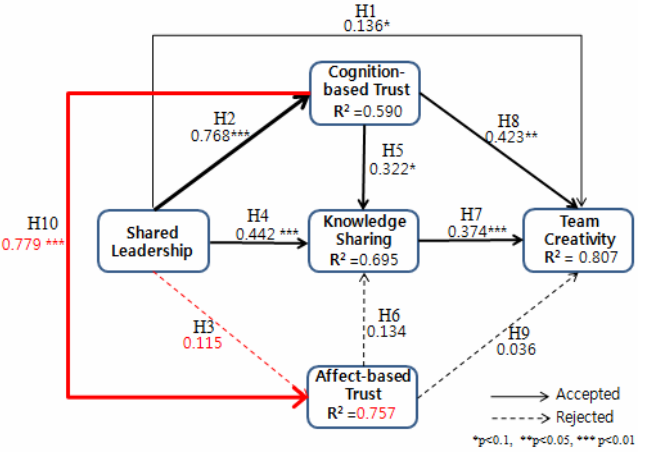


Fig. 4. Revised PLS Model

4 Discussion

4.1 Theoretical and Practical Implications

This study's objective was to intensify the theory of team creativity by examining the effects of shared leadership, knowledge-sharing, and interpersonal trust on team creativity using a team-based approach. As for theoretical importance, we selected and tested constructs that are most closely related to creativity at the team level. Above all, we intensified our PLS model by using Bayesian network (BN) approach as an ancillary role. Unexpected results from the results of BN helped us to revise the initial PLS model. We constructed a new PLS model through an additional hypothesis and interpretations. An empirical study for our integrative team creativity model is presented to demonstrate the successful application of the proposed method. Therefore, we suggest that many researchers employ our Bayesian network approach.

This study has important implications for team leaders and managers. Given the need for team creativity in solving complex challenges faced by organizations, managers should ensure that each team has a clear and shared sense of direction and purpose, thereby promoting participation in team activities and identifying the organizational contexts in which shared leadership, knowledge-sharing, and cognition-based trust are most likely to enhance team creativity.

4.2 Limitations and Future Research

This paper has limitations that should be addressed in future research. First, our study targeted student samples in an e-learning course, which is a non-face-to-face

environment. Consequently, we did not confirm the effects of affect-based trust. Because the samples were students, not full-time employees, the results may differ in an actual workplace. Future research should pursue comparative studies in various environments. Second, an important focus for future research is the long-term effects (i.e., whether team members' reactions to constructs were temporary or whether such reactions were permanent) of shared leadership, knowledge-sharing and interpersonal trust. Third, we did not consider many important antecedents of team creativity. Future research should probe into how personal traits (such as age, level of education, and working experience), diversity within a team, and organizational characteristics (such as firm size and industry type) may moderate the relationships among the constructs. Finally, we didn't apply various algorithms (such as hill climbing, K2, simulated annealing, genetic, tabu and so on) of Bayesian networks to the proposed approach. Moreover, we didn't fully explain indirect effects in the DAG of the Bayesian network. In our results, shared leadership indirectly contributes to knowledge sharing through cognition-based trust. Future research should implement other BN algorithms, and examine theoretical backgrounds for the explanation of indirect influence.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Ajzen, I., Fishbein, M.: Attitudinal and Normative Variables as Predictors of Specific Behavior. Journal of Personality and Social Psychology 27(1), 41–57 (1973)
2. Amabile, T.M.: A Model of Creativity and Innovation in Organizations. In: Staw, B.M., Cummings, L.L. (eds.) Research in Organizational Behavior, vol. 10, pp. 123–167. JAI Press, Greenwich (1988)
3. Anderson, R.D., Vastag, G.: Causal Modeling Alternatives in Operations Research: Overview and Application. European Journal of Operational Research 156(1), 92–109 (2004)
4. Argote, L.: Organizational learning: Creating, Retaining, and Transferring Knowledge. Kluwer Academic, Boston (1999)
5. Avolio, B.J., Jung, D.I., Sivasubramaniam, N.: Building Highly Developed Teams: Focusing on Shared Leadership Processes, Efficacy, Trust, and Performance. In: Beyerlein, M.M., Johnson, D.A. (eds.) Advances in Interdisciplinary Study of Work Teams: Team Leadership, vol. 3, pp. 173–209. JAI Press, Greenwich (1996)
6. Bidault, F., Castello, A.: Trust and Creativity: Understanding the Role of Trust in Creativity-oriented Joint Developments. R&D Management 39(3), 259–270 (2009)
7. Carson, J.B., Tesluk, P.E., Marrone, J.A.: Shared Leadership in Teams: An Investigation of Antecedent Conditions. Academy of Management Journal 50(5), 1217–1234 (2007)
8. Cerquides, J., De Mantaras, R.L.: TAN Classifiers Based on Decomposable Distributions. Machine Learning 59, 323–354 (2005)
9. Chen, M.H., Chang, Y.C., Hung, S.C.: Social Capital and Creativity in R&D Project Teams. R&D Management 38(1), 21–34 (2008)
10. Chow, C.K., Liu, C.N.: Approximating Discrete Probability Distributions with Dependence trees. IEEE Transactions on Information Theory 14(3), 462–467 (1968)

11. Chowdhury, S.: The Role of Affect- and Cognition-based Trust in Complex Knowledge Sharing. Journal of Managerial Issues 17(3), 310–326 (2005)
12. Chua, R.Y.J., Morris, M.W.: From the Head and the Heart: Locating Cognition-and Affect-based Trust in Managers' Professional Networks. Academy of Management Journal 51(3), 436–452 (2008)
13. Day, D.V., Gronn, P., Salas, E.: Leadership Capacity in Teams. The Leadership Quarterly 15(6), 857–880 (2004)
14. De Clercq, D., Thongpapanl, N.T., Dimov, D.: The Role of Conflict and Social Capital in Cross-Functional Collaboration. In: Proceedings of the 4th Workshop on Trust within and Between Organizations, Amsterdam (October 2007)
15. De Vries, R.E., van den Hooff, B., de Ridder, J.A.: Explaining Knowledge Sharing: The Role of Team Communication Styles, Job Satisfaction, and Performance Beliefs. Communication Research 33(2), 115–135 (2006)
16. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning 29(2-3), 131–163 (1997)
17. Jehn, K.A.: A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict. Administrative Science Quarterly 40, 256–282 (1995)
18. Johnson, D., Grayson, K.: Cognitive and Affective Trust in Service Relationships. Journal of Business Research 58, 500–507 (2005)
19. Lapierre, J., Giroux, V.: Creativity and Work Environment in a High-Tech Context. Creativity and Innovation Management 12, 11–23 (2003)
20. Lauria, E.J.M., Duchessi Peter, J.: A Methodology for Developing Bayesian Networks: An Application to Information Technology (IT) Implementation. European Journal of Operational Research 179, 234–252 (2007)
21. Lauritzen, S.L.: Causal Inference in Graphical Models. In: Barndorff-Nielsen, O.E., Cox, D.R., Kluppelberg, C. (eds.) Complex Stochastic Systems. Chapman & Hall, London (2000)
22. Lewis, J.D., Weigert, A.: Trust as a Social Reality. Soc. Forces 1985 63, 967–985 (1985)
23. Mamykina, L., Candy, L., Edmonds, E.: Collaborative Creativity. Communications of the ACM 45(10) (2002)
24. Mayo, M., Meindl, J.R., Pastor, J.C.: Shared Leadership in Work Teams: A Social Network Approach. In: Pearce, C.L., Conger, J.A. (eds.) Shared Leadership: Reframing the Hows and Whys of Leadership, pp. 193–214. Sage, Thousand Oaks (2003)
25. McAllister, D.J.: Affect-and Cognition-based Trust as Foundations for Interpersonal Cooperation in Organizations. Academy of Management Journal 38, 24–59 (1995)
26. Mehra, A., Smith, B., Dixon, A., Robertson, B.: Distributed Leadership in Teams: The Network of Leadership Perceptions and Team Performance. Leadership Quarterly 17, 232–245 (2006)
27. Mohrman, S.A., Cohen, S.G., Mohrman, A.M.: Designing Team-based Organizations: New Forms for Knowledge Work. Lavoisier (1995)
28. Mooradian, T., Renzl, B., Matzler, K.: Who Trusts? Personality, Trust and Knowledge Sharing. Management Learning 37(4), 523–540 (2006)
29. Oldham, G.R.: Stimulating and Supporting Creativity in Organizations. In: Jackson, S.E., Hitt, M.A., DeNisi, A.S. (eds.) Managing Knowledge for Sustained Competitive Advantage: Designing Strategies for Effective Human Resource Management, pp. 243–273. Jossey-Bass, San Francisco (2003)
30. Oldham, G.R., Cummings, A.: Employee Creativity: Personal and Contextual Factors at Work. Academy of Management Journal 39, 607–634 (1996)

31. Organ, D.W.: The Motivational Basis of Organizational Citizenship Behavior. In: Staw, B.M., Cummings, L.L. (eds.) Research in Organizational Behavior, vol. 12, pp. 43–72. JAI Press, Greenwhich (1990)
32. Pearce, C.L., Sims, H.P.: The Relative Influence of Vertical vs. Shared Leadership on the Longitudinal Effectiveness of Change Management Teams. Group Dynamics: Theory, Research, and Practice 6(2), 172–197 (2002)
33. Podsakoff, P.M., MacKenzie, S.B., Moorman, R.H., Fetter, R.: Transformational Leader Behaviours and Their Effects on Followers' Trust in Leader, Satisfaction, and Organizational Citizenship Behaviors. Leadership Quarterly 1, 107–142 (1990)
34. Robinson, S.L.: Trust and Breach of the Psychological Contract. Administrative Science Quarterly 41(4), 574–599 (1996)
35. Shin, S.J., Zhou, J.: Transformational Leadership, Conservation, and Creativity: Evidence from Korea. Academy of Management Journal 46(6), 703–714 (2003)
36. Simons, T., Peterson, R.: Task Conflict and Relationship Conflict in Top Management Teams: the Pivotal Role of Intragroup Trust. Journal of Applied Psychology 85, 102–111 (2000)
37. Wu, W.W.: Linking Bayesian Networks and PLS Path Modeling for Causal Analysis. Expert Systems with Applications 37, 134–139 (2010)
38. Wu, W.L., Hsu, B.F., Yeh, R.S.: Fostering the Determinants of Knowledge Transfer: A Team-level Analysis. Journal of Information Science 33(3), 326–339 (2007)
39. Yukl, G.: Leadership in Organizations. Prentice-Hall, Upper Saddle River (2002)

Effects of Users' Perceived Loneliness and Stress on Online Game Loyalty

Bong-Won Park[1] and Kun Chang Lee[2,*]

[1] Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
combio00@naver.com
[2] Professor of MIS at SKK Business School
WCU Professor at Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
Tel.: +82 7600505; Fax: +82 2 7600440
kunchanglee@gmail.com

Abstract. Online games become very popular as the Internet permeates down to all ranks of our lives. However, contrary to our common sense that those users loyal to the online games will just enjoy the games only for the sake of pleasure, user's perceived loneliness and stress are believed to affect their loyalty for the games. However, this research issue remains unexplored sufficiently in the fields of IS studies. In this sense, this paper proposes a new research model in which users' perceived loneliness and stress have relationships to game loyalty through other experiential factors such as flow, enjoyment, and character identification. To prove the validity of our proposed research model, empirical analysis was performed with 187 valid questionnaires using PLS (Partial Least Square). Results revealed that the proposed research model is statistically significant, and loneliness and perceive stress hold crucial position in the users' loyalty to games.

Keywords: online game, perceived stress, loneliness, character identification, enjoyment, flow.

1 Introduction

The Internet has become a very integral part of modern daily life. Due to the availability of high speed Internet at locations across the globe, previously impossible things have become possible due to the Internet. Specifically, we can transfer money to others via electronic banking while listening to music via online radio. In addition, local Internet users can join a specific online game and play with global gamers.

As many people consider online gaming a leisure, and as young children are becoming increasingly familiar with the computer and the Internet, the market volume

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 556–565, 2011.

of this online gaming industry and the number of online gamers are increasing, with the global revenue of online gaming expected to reach $24.8 billion by 2013 [1]. This increase in revenue creates new jobs such as professional gaming and new markets such as Internet cafés. In addition, game-based learning is a new topic in the educational sphere. However, the development of game addiction has resulted in many negative by products such divorce, job loss and school absences.

To help analyze the positive and negative effects of online gaming, game development companies and researchers are now studying the motivations behind online game play. However, most research focuses on only one set of aspects, either the psychological or the experiential. In addition, much research focuses on young students and adolescents. However, as the number of older people playing online games has been increasing and because many years have passed since the introduction of online games [2], a broad survey including older game users is needed [3].

In this study, we aimed to determine which characteristics affect loyalty for online gaming. Specifically, the psychological factors, including loneliness and perceived stress, and experiential factors, including character identification, flow, and perceived enjoyment, were studied. In addition, our data included individuals greater than forty years of age.

2 Previous Studies

Studies regarding online games can be broadly classified into three categories. The first is adoption, which includes the characteristics that affect game playing. The second is addiction, when a game user feels a compulsion to use the online game. The final category is the business model of online games such as a market for game items.

2.1 Adoption of Online Games

By using and extending the Technology Adoption Model (TAM) or the Theory of Reasoned Action (TRA), many researchers endeavor to explain why users play online games. For example, Hsu and Lu [4] found that the flow experience is an important factor in the intention to play an online game. Wu and Liu [5] uncovered that trust in online game websites and online game enjoyment are positively related to the attitude toward online games. In addition, Wang and Wang [6] found that male game users have a higher intention to play online game than do female game users. Furthermore, as increasingly older game users play online games, Williams et al. [3] insisted on the need for a broad sampling that is not focused only on young people.

2.2 Addictions to Online Games

As the Internet becomes more popular and is used by more people, Internet addiction is becoming more prevalent. Young [7] defined Internet addiction as "an impulse-control disorder which does not involve an intoxicant" and developed a questionnaire to measure Internet addiction. This scale is used worldwide to measure Internet addiction [8, 9, 10]. In addition, this measure can also be applied to online game addiction after modifying some of the meanings [11, 12].

2.3 Business Model of Online Games

The online game business model is classified into a subscription model and a free-to-play model. In the subscription model, in order to play a specific game, the game users subscribe to that game on a monthly basis. However, in the free-to-play model, a person can play the online game free of charge. However, to adorn or increase the power of one's game character (i.e., avatar), online game items (virtual products) must be purchased. Therefore, Lin and Sun [13] insisted that these game users are no longer simple players but have also become consumers. The motivations for purchasing game items are the perceived playfulness, character competency, and the requirements of the quest system [14]. In addition, social networking services such *Facebook* and *MySpace* now include social games in which users can purchase game items to enhance their game environments.

3 Hypothesis Development

3.1 Loneliness, Perceived Stress and Character Identification

Loneliness is defined as an unpleasant response experienced by a person who has not established close and meaningful relationships with others [15]. Modern society may foster these feelings of loneliness, leading to a population which contains many lonely individuals.

People experience stress from many sources, including work, school, and their personal lives. Through interactions with other, however, people come to realize that these stresses are common, and this knowledge of shared experience can help to alleviate some of the stress. However, lacking these sorts of social interactions, lonely people will experience enhanced stress. Previous research has shown that loneliness is moderately correlated with perceived stress [16, 17]. Hence, we formed the following hypothesis:

Hypothesis 1a: Loneliness is positively related to perceived stress.

In online games, users create game characters (i.e., avatars). When gamers play online games continuously, they may purchase game items to adorn a game character or to increase its power. Therefore, as the cumulative time of playing an online game increases, the user begins to relate to the game character, coming to view the avatar as an online version of her/himself. For example, if the game character is killed by others in an online game, the gamers feel that they have been attacked by others in the real world. This characteristic can be defined as character identification, and researchers have confirmed that game users develop a close relationship with their game characters [18, 19].

Lonely individuals often use the Internet to connect with other people [20, 21]. In addition, they may express themselves better online than they are able to do in an offline environment. Therefore, these lonely users more easily become attached to online games and their game characters [22, 23]. Based on these facts, we present the following hypothesis:

Hypothesis 1b: Loneliness is positively related to character identification.

3.2 Perceived Stress, Character Identification, Enjoyment and Flow

Online games are often used as a means to escape from stress. Since online games have no definitive ending, game users may invest unintended amounts of time into their game playing. During online game play, uses think less about their current problems, focusing instead on the game itself. Therefore, a stressed person is able to more easily experience the flow of the game compared to the experience of others who are not stressed. Henceforth, the following hypothesis is presented.

Hypothesis 2a: Perceived stress is positively related to the flow.

When playing online games, many gamers do often become unaware of the current time and experience a rapid passage of time. This characteristic is defined as flow, "the holistic sensation that people feel when they act with total involvement" [24]. This flow experience may develop and increase when game users interact with their game characters. The more a user plays a game, the more identity he/she will feel with the game character and the more attention he/she will invest in the game. Therefore, the experience of flow increases with character identification. Therefore, the following hypothesis is derived.

Hypothesis 2b: Character identification is positively related to the flow.

The feelings of game users often become dependent on their success or failure in the game. These feelings are often impacted by the strength of the relationship between the game user and the character. When a game user identifies with a game character, this feeling will increase. However, if s/he does not identify with the game character, this feeling will decrease. In accordance with this concept, we propose the following hypothesis.

Hypothesis 2c: Character identification is positively related to perceived enjoyment.

Many people engage in online gaming to experience fun and to escape the stress of daily life [25, 26]. Generally, if a TV program is funny, a person does not want to stop watching. Similarly, when online gaming provides pleasure, the user wants to continuously play. While playing that game, the user is unaware of the current time and of how much time has elapsed during game play. Therefore, enjoyment will increase the possibility of flow. In addition, the enjoyment experienced from playing a game is one of the factors that affect the flow experience [27]. Hence, we form the following hypothesis:

Hypothesis 2d: Perceived enjoyment is positively related to flow.

3.3 Flow and Loyalty

On Internet sites, the flow experience is positively related to the loyalty to that website, such as those for online tours [28] or professional sports [29].

When game users experience flow, they are able to forget their current problems and stress, thus increasing the desire to play. In addition, users recommend the game to others based on the positive flow experience [30]. Using these facts, we present the following hypothesis:

Hypothesis 3: Flow is positively related to online game loyalty.

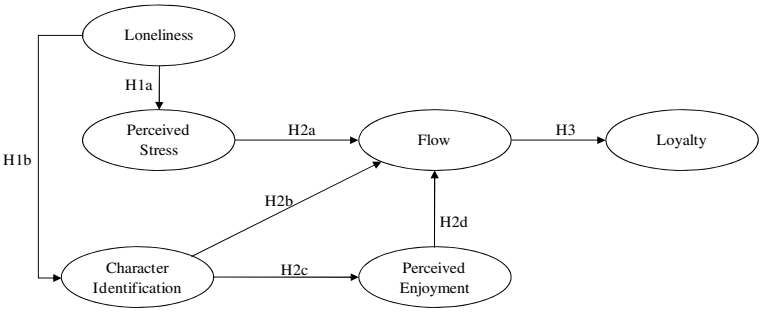


Fig. 1. Research Model

4 Empirical Analysis

4.1 Questionnaire Survey and Sample Statistics

The survey items that we used were adopted from the literature. First, to measure loneliness, we used a short-form UCLA Loneliness Scale (ULS-8) [31] consisting of eight items. This scale is as reliable and valid as is the long form UCLA loneliness scale. Second, to measure the perceived stress, we used the four-item Perceived Stress Scale (PSS) [32, 33]. Third, to measure character identification, we used questionnaires from two previous studies of Hefner et al. and Fornell and Larcker [18, 34] and additional questionnaires modified from famous literature. All of the questionnaires used a seven-point scale, ranging from completely disagree to completely agree.

The samples were collected using an online survey in Korea. The respondents ranged from teenagers to users in their fifties (10-19: 29, 20-29: 36, 30-39: 45, 40-49: 43, and 50-59: 34). Almost the same number of male and female respondents participated in this survey, and there were 187 total respondents.

4.2 Reliability and Confirmatory Factor Analyses

Preliminary analysis

Six constructs of loneliness, loyalty, character identification, flow, enjoyment, and perceived stress were used in this analysis. The survey items of each construct were reliable, with Cronbach's alpha values greater than 0.7. The validities of the survey items were tested using a principal component analysis with varimax rotation.

Table 1. Results of reliability and factor analyses

Survey item	Cronbach's α	LONE	LOYA	CHAR	FLOW	ENJ	STR
loneliness5	0.934	0.883					
loneliness1		0.870					
loneliness2		0.868					
loneliness8		0.839					
loneliness4		0.832					
loneliness7		0.806					
loneliness6		0.724					
loyalty4	0.893		0.844				
loyalty5			0.770				
loyalty1			0.762				
loyalty3			0.762				
loyalty2			0.754				
char_identi2	0.913			0.882			
char_identi4				0.861			
char_identi1				0.842			
char_identi3				0.794			
game_flow1	0.905				0.840		
game_flow3					0.821		
game_flow2					0.806		
game_flow4					0.735		
game_enjoyment2	0.887					0.812	
game_enjoyment3						0.779	
game_enjoyment1						0.766	
game_enjoyment4						0.656	
perceived_stress3	0.745						0.827
perceived_stress2							0.744
perceived_stress4							0.709
perceived_stress1							0.629
Eigenvalue		7.485	6.171	2.448	2.071	1.348	1.115
Variance explained		26.732	22.040	8.743	7.397	4.816	3.981
Total variance explained (%)		26.732	48.772	57.515	64.912	69.727	73.709

Note 1: LONE: Loneliness, LOYA: Loyalty, CHAR: Character Identification, FLOW: Flow, ENJ: Enjoyment, STR: Perceived Stress

Note 2: One item of loneliness (item #3) was not included in the factor analysis.

The first factor, loneliness, explained 26.73% of the total variance; the second factor, loyalty, accounted for 22.04%. The total variance explained by the six factors was 73.7%, as shown in Table 1. From these results, we concluded that the survey items were statistically valid.

The measurement model

To confirm the reliability and validity of the measurement data, we performed a confirmatory factor analysis. In the reliability test, all of the composite reliabilities were greater than 0.7, and the AVEs (Average Variance Extracted) were greater than 0.5 [35]. For the validity test, the correlation between two factors was less than the square root of the AVE value of each factor [35]. Therefore, as shown in Table 2, these data were reliable and valid.

Table 2. Correlations of latent variables and AVEs

Construct	Composite Reliability	LONE	LOYA	CHAR	FLOW	ENJ	STR
Loneliness	0.946	0.847					
Loyalty	0.921	-0.103	0.837				
Character Identification	0.939	0.242	0.283	0.891			
Flow	0.934	0.138	0.434	0.496	0.883		
Enjoyment	0.920	-0.170	0.665	0.243	0.490	0.862	
Perceived Stress	0.830	0.321	-0.046	0.221	0.289	-0.010	0.743

Note: Values on the underlined diagonal are the square roots of the AVEs.

The structural model

This study's hypotheses were tested using SmartPLS 2.0 with the bootstrapping procedure, and all of the hypotheses were accepted at a 99% confidence level.

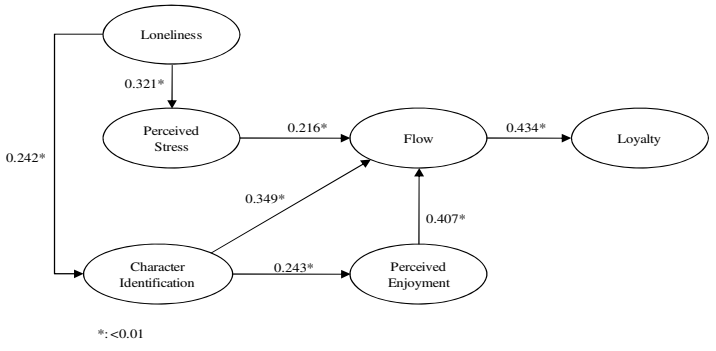


Fig. 2. Results of the structural model

4.3 Discussion

This paper addressed the psychological and experiential factors that affect customer loyalty in online games using the PLS (Partial Least Square). From this analysis, we reached the following conclusions. First, lonely people experience greater stress because they do not have people to talk with or with whom to relieve stress. In addition, if a game user feels lonely, s/he tends to attach to her/his game character through character identity, and s/he plays the online game to experience interactions with others.

Second, a person who is feeling stress tends to more easily identify with the game character to experience enjoyment from the online game, increasing the state of flow. While playing the game, a gamer forgets his/her current problems and focuses only on playing the game. Therefore, the player reaches the state of flow. In addition, if the gamer identifies with the game character, s/he focuses more intently on the game. Therefore, character identification positively impacts the flow of the game.

Third, in an online game, when a gamer's character kills monsters or defeats other players, the player is happy. Similarly, if a person identifies with the game character, s/he finds the online game enjoyable. Therefore, character identification is an important factor in enhancing online gaming enjoyment.

Finally, a game user who experiences flow will want to play the online game again and will recommend the game to other people. Our results confirm the results of previous research, and game development companies should develop online games to ensure the feeling of flow during game play.

5 Concluding Remarks

Advances in high speed Internet have helped the online gaming industry flourish. Without meeting in the real world, users can meet other individuals in the cyber world (i.e., via an online game). By playing online games, game users relieve stress and experience enjoyment. Many years have passed since online games were introduced, contributing to the variety of ages that now play online games. However, there has been limited research considering both the psychological and experiential factors, as well as limited studies of users from various age groups.

In this study, we aimed to identify the factors related to loyalty in online gaming in users of various ages. Using PLS analysis, we found that loneliness is positively related to perceived stress and character identification in gaming. Second, perceived stress, character identification, and enjoyment are positively related to gaming flow. Third, character identification is positively related to enjoyment. Finally, the flow is positively related to loyalty to an online game.

These results help us to understand online game users and should be considered by online game developers and publishers in order to increase the market sizes of online games. This study reports an approach which considers both psychological and experiential factors in online gaming. However, many factors may also depend on the game users' favorite game genres and ages. These factors will be analyzed in detail in future works.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Wu, J.: Global Video Game Market Forecast, Strategy Analytics (2010)
2. Abraham, L.B., Mörn, M.P., Vollman, A.: Women on the Web: How Women are Shaping the Internet, comScore (2010)
3. Williams, D., Yee, N., Caplan, S.E.: Who Plays, How Much, and Why? Debunking the Stereotypical Gamer Profile. Journal of Computer-Mediated Communication 13, 993–1018 (2008)
4. Hsu, C.L., Lu, H.P.: Why do People Play On-Line Games? An Extended TAM with Social Influences and Flow Experience. Information & Management 41, 853–868 (2004)

5. Wu, J., Liu, D.: The Effects of Trust and Enjoyment on Intention to Play Online Games. Journal of Electronic Commerce Research 8, 128–140 (2007)
6. Wang, H.Y., Wang, Y.S.: Gender Differences in the Perception and Acceptance of Online Games. British Journal of Educational Technology 39, 787–806 (2008)
7. Young, K.: Caught in the Net. John Wiley & Sons, Chichester (1998)
8. Kim, K., Ryu, E., Chon, M., Yeun, E., Choi, S., Seo, J., Nam, B.: Internet Addiction in Korean Adolescents and its Relation to Depression and Suicidal Ideation: A Questionnaire Survey. International Journal of Nursing Studies 43, 185–192 (2006)
9. Leung, L.: Net-Generation Attributes and Seductive Properties of the Internet as Predictors of Online Activities and Internet Addiction. Cyberpsychology & Behavior 7, 333–348 (2004)
10. Yoo, H., Cho, S., Ha, J., Yune, S., Kim, S., Hwang, J., Chung, A., Sung, Y., Lyoo, A.: Attention Deficit Hyperactivity Symptoms and Internet Addiction. Psychiatry and Clinical Neurosciences 58, 487–494 (2004)
11. Johansson, A., Götestam, K.G.: Problems with Computer Games without Monetary Reward: Similarity to Pathological Gambling. Psychological Reports 95, 641–650 (2004)
12. Lu, H.-P., Wang, S.-m.: The Role of Internet Addiction in Online Game Loyalty: An Exploratory Study. Internet Research 18, 499–519 (2008)
13. Lin, H., Sun, C.T.: Cash Trade within the Magic Circle: Free-to-play Game Challenges and Massively Multiplayer Online Game Player Responses. In: DiGRA 2007: Situated Play, pp. 335–343 (2007)
14. Guo, Y., Barnes, S.: Virtual Item Purchase Behavior in Virtual Worlds: an Exploratory Investigation. Electronic Commerce Research 9, 77–96 (2009)
15. Russell, D.: The Measurement of Loneliness. In: Peplau, L.A., Perlman, D. (eds.) Loneliness: A Sourcebook of Current Theory, Research and Therapy, pp. 81–104. John Wiley and Sons, New York (1982)
16. Cacioppo, J.T., Hughes, M.E., Waite, L.J., Hawkley, L.C., Thisted, R.A.: Loneliness as a Specific Risk Factor for Depressive Symptoms: Cross-Sectional and Longitudinal Analyses. Psychology and Aging 21, 140–151 (2006)
17. Miczo, N.: Humor Ability, Unwillingness to Communicate, Loneliness, and Perceived Stress:Testing a Security Theory. Communication Studies 55, 209–226 (2004)
18. Hefner, D., Klimmt, C., Vorderer, P.: Identification with the Player Character as Determinant of Video Game Enjoyment. In: Ma, L., Rauterberg, M., Nakatsu, R. (eds.) ICEC 2007. LNCS, vol. 4740, pp. 39–48. Springer, Heidelberg (2007)
19. McDonald, D.G., Kim, H.: When I Die, I Feel Small: Electronic Game Characters and the Social Self. Journal of Broadcasting & Electronic Media 45, 241–258 (2001)
20. Davis, R.A.: A Cognitive-Behavioral Model of Pathological Internet Use. Computers in Human Behavior 17, 187–195 (2001)
21. Kim, J., LaRose, R., Peng, W.: Loneliness as the Cause and the Effect of Problematic Internet Use: The Relationship between Internet Use and Psychological Well-being. CyberPsychology & Behavior 12, 451–455 (2009)
22. Griffiths, M.: Violent Video Games and Aggression: A Review of the Literature. Aggression and Violent Behavior 4, 203–212 (1999)
23. McKenna, K.Y.A., Green, A.S., Gleason, M.E.J.: Relationship Formation on the Internet: What's the Big Attraction? Journal of Social Issues 58, 9–31 (2002)
24. Csikszentmihalyi, M.: Beyond Boredom and Anxiety. Jossey-Bass, San Francisco (1975)
25. Wan, C., Chiou, W.: Why are Adolescents Addicted to Online Gaming? An Interview Study in Taiwan. Cyber Psychology & Behavior 9, 762–766 (2006)

26. Yee, N.: Motivations for Play in Online Games. Cyber Psychology & Behavior 9, 772–775 (2006)
27. Ha, I., Yoon, Y., Choi, M.: Determinants of Adoption of Mobile Games under Mobile Broadband Wireless Access Environment. Information & Management 44, 276–286 (2007)
28. Wu, J.J., Chang, Y.S.: Towards Understanding Members' Interactivity, Trust, and Flow in Online Community. Industrial Management & Data Systems 105, 937–954 (2005)
29. O'Cass, A., Carlson, J.: Examining the Effects of Website-induced Flow in Professional Sporting Team Websites. Internet Research 20, 115–134 (2010)
30. Lee, S.-C., Xiang, J.-Y., Gu, J.C., Suh, Y.-H.: Determinants of Effecting Customer Loyalty: Comparison among Korean, Japanese and Chinese Online Game Market. Korea Journal of Management Science 23, 41–57 (2006)
31. Hays, R.D., DiMatteo, M.R.: A Short-form Measure of Loneliness. Journal of Personality Assessment 51, 69–81 (1987)
32. Cohen, S., Kamarck, T., Mermelstein, R.: A Global Measure of Perceived Stress. Journal of Health and Social Behavior 24, 385–396 (1983)
33. Cohen, S., Williamson, G.: Perceived Stress in a Probability Sample of the U.S. In: Spacapam, S., Oskamp, S. (eds.) The Social Psychology of Health: Claremont Symposium on Applied Social Psychology. Sage, Newbury Park (1988)
34. Cohen, J.: Defining Identification: A Theoretical Look at the Identification of Audiences with Media Characters. Mass Communication & Society 4, 245–264 (2001)
35. Fornell, C., Larcker, D.: Structural Equation Models with Unobservable Variables and Measurement Error. Journal of Marketing Research 18, 39–50 (1981)

An Adjusted Simulated Annealing Approach to Particle Swarm Optimization: Empirical Performance in Decision Making

Dae Sung Lee[1], Young Wook Seo[2,*], and Kun Chang Lee[3,*]

[1] PhD Candidate at SKK Business School, Sungkyunkwan University
Seoul 110-745, Republic of Korea
leeds1122@gmail.com
[2] Researcher, Software Engineering Center at NIPA, Seoul 138-711, Republic of Korea
Tel.: +82 2 760 0505; Fax: +82 2 760 0440
seoyy123@gmail.com
[3] Professor at SKK Business School
WCU Professor at Department of Interaction Science
Sungkyunkwan University, Seoul 110-745, Republic of Korea
Tel.: +82 2 760 0505; Fax: +82 2 760 0440
kunchanglee@gmail.com

Abstract. Particle swarm optimization (PSO) is a novel population-based searching technique proposed as an alternative to genetic algorithm (GA). It has had wide applications in a variety of fields. We suggest a hybrid clustering algorithm, which applies the combination of conventional PSO and SA (Simulated Annealing) algorithm to the process of K-means clustering in order to solve the problem of premature convergence. In addition we develop an adjustment algorithm, which modifies the acceleration constants of PSO by comparison of global and local best position, and is applied to the mixture algorithm named as SA-PSO so as to minimize the search of unnecessary areas and enhance performance. We simulated and compared three algorithms (K-PSO, SA-PSO and Adjusted SA-PSO). The results demonstrated our new approach (Adjusted SA-PSO) had the most excellent performance in usefulness and reliability evaluation, which denotes fitness function and mean absolute error respectively.

Keywords: K-means clustering, PSO (Particle Swarm Optimization), SA algorithm (Simulated Annealing), SA-PSO, Adjusted SA-PSO.

1 Introduction

The study of nonlinear functions, which have local and global solutions, has been an interesting subject for many scientists to delve into. In search of optimization solution, one of the most effective and useful methods is probabilistic optimization. Evolutionary computation (EC) is a probabilistic optimization algorithm to provide a

* Corresponding authors.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6592, pp. 566–575, 2011.

valuable solution for clustering. There have been various studies using clustering-based evolutionary computation techniques such as Genetic Algorithm (GA), Evolution Strategies (ES), Particle Swarm Optimization (PSO) and so on [1, 3, 14].

Particle Swarm Optimization (PSO) is a type of EC techniques that was first proposed by Kenney and Eberhart in 1995 [8]. PSO was inspired by the social behavior of organisms such as bird flocking or fish schooling. PSO algorithm is easy to implement and takes a short time for calculation, and does not need a large memory. However, it still has the problem of premature convergence [12, 14]. To solve the weakness, many studies have introduced SA (Simulated Annealing) into PSO algorithm, which are named as SA-PSO [6, 7, 17]. Notwithstanding the performance of SA-PSO, the jump attribute of the SA bring about another premature convergence. In this respect, it requires the continuous development of PSO-based algorithm.

In this paper, we try to apply hybrid K-means clustering and PSO-based algorithm to carry out data clustering with Matlab tool. To find the most effective clustering method, we suggest a new approach named as "Adjusted SA-PSO algorithm", which would improve the hybrid algorithm of PSO and SA. Adjusted SA-PSO modifies the existing SA-PSO algorithm by adding functions to prevent from falling into local optimization solution. We compare the proposed approach with existing clustering algorithms and evaluate the usefulness and reliability.

2 Algorithms to Cluster Data

2.1 K-Means Clustering

K-means clustering [13] groups data vectors into a predefined number of clusters, based on Euclidean distance as similarity measure. The K-means method is a widely used clustering procedure that searches for a nearly optimal partition with a fixed number of clusters. Data vectors within a cluster have small Euclidean distances from one another, and are associated with one centroid vector. The centroid vector is the mean of the data vectors that belong to the corresponding cluster. The process of K-means clustering is the equation (1) and (2) in Section 3 below. The K-means algorithm has been popular because of its easiness and simplicity for application. However, above all it may converge to a local minimum under certain conditions. To eliminate the premature convergence, The K-means has widely been combined with other algorithms.

2.2 PSO (Particle Swarm Optimization)

Particle swarm optimization (PSO) is an evolutionary optimization technique developed by Kennedy and Eberhart [8]. PSO has been known to be a powerful tool to solve problems characterized by nonlinearity, multi-optimization, and multi-dimensionality through adaptation derived from the theory of social psychology. In PSO algorithm, each particle has a simple individual behavior that results in a

complex emergent global performance. Basically, each individual analyses its current state comparing with its own experience and the experience of others. The goal of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function. Each particle represents a position in an N dimensional space, and is flown through this multi-dimensional search space, adjusting its position toward both the particle's best position found thus far and the best position in the neighborhood of that particle [16]. Each particle is adjusted by the equations (9) and (10) in the section 3 below. In equation (9), w is the inertia weight that controls the convergence of the particles; Y_i is the vector containing the best position of particle i; Y_i is the vector containing the best position among all particles within a pre-defined neighborhood; C_1 is the stochastic weight vector that will weigh the influence of the cognitive component; C_1 and C_2 is the weight vector that will weigh the influence of the social component. Van den Bergh [15] demonstrated that when $w = 0.72$ and $C_1 = C_2 = 1.49$ these values ensured good convergence. However, we initialize the specific value of C_1 and C_2, and introduce their adjustment into our proposed method in the section 3.

2.3 SA (Simulated Annealing)

In its original form [9], SA (Simulated Annealing) algorithm is based on the analogy between the simulation of the annealing of solids and the problem of solving large combinatorial optimization problems [10]. For this reason the algorithm became known as "simulated annealing". SA improves the shortcoming of conventional heuristic techniques based on repeated improvements which converge into local minimum solution. In condensed matter physics, annealing denotes a physical process in which a solid in a heat bath is heated up by increasing the temperature of the heat bath to a maximum value at which all particles of the solid randomly arrange themselves in the liquid phase, followed by cooling through slowly lowering the temperature of the heat bath [10]. In this way, all particles arrange themselves in the low energy ground state of a corresponding lattice, provided the maximum temperature is sufficiently high and the cooling is carried out sufficiently slowly. The cooling phase of the annealing process starts off from a randomly selected point within the search space. If the fitness of a new candidate solution is less than the fitness of the current solution, the new candidate solution is not automatically rejected. Instead it becomes the current solution with a certain transition probability $p(T)$. If ΔE is the difference between previous candidate solution and current one, this transition probability depends on the difference in the fitness ΔE and the temperature T. Here, 'temperature' is an abstract control parameter for the algorithm rather than a real physical measure. In Section 3 below, the equation (5) includes a common transition function $p(T) = exp(-\Delta E /T)$ for a given temperature and a given difference in fitness. The algorithm starts with a high temperature, which is subsequently reduced slowly, usually in steps. In Section 3 below, the equation (7) shows the standard cooling function introduced by Kirkpatrick. In that equation, T_n is temperature at step n, α is cooling coefficient ($\alpha < 1$). Many others can be found in the literature [4, 5].

3 Adjusted SA-PSO

In order to further improve the space exploration capability of the particles, we introduced variation factors to our new approach. The equation (8) below represents an added algorithm that we applies to a mixture of SA and PSO in the process of updating *P*best and *G*best. As shown in Fig. 1, we intend to phase in an individual algorithm that is added on hybrid algorithms one by one.

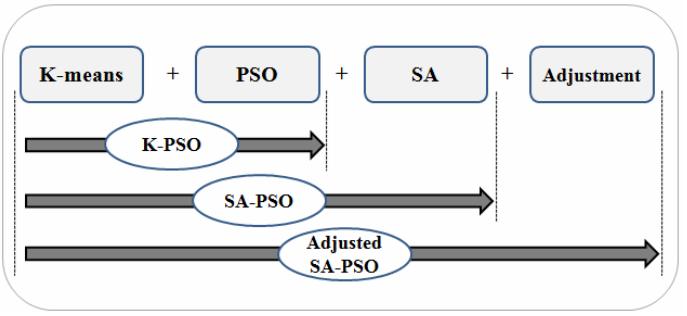


Fig. 1. Flow of Algorithms

3.1 SA-PSO for Clustering

Many studies have tried to combine PSO and SA algorithm to improve the performance of heuristic optimization technique. Da and Xiurun [2] proposed SAPSO algorithm which added the jump attribute of SA to the inertia weight of PSO, and used it to find the weight of Artificial Neural Network (ANN). PSOSA algorithm, proposed by Liang, et al. [11], is basically similar to the SAPSO algorithm in that the jump attribute of SA is applied in the process of updating the velocity of the *G*best particles of PSO. By employing the SA-based selection for the best position when updating the velocity in PSO, the hybrid strategy is of more effective global exploration ability over pure PSO at the beginning searching stage (when temperature is high) so as to avoid premature convergence. As the temperature decreases, the hybrid strategy transforms to PSO smoothly to stress the exploitation.

In this paper, we apply the jump property of SA to the process of updating the global best and local best positions in PSO clustering. This updating process requires the function f which represents the mean of Euclidean distance. If the f value of the next position ($f(X_i(t+1))$) is greater (worse) than that of the current *P*best ($f(Y_i(t))$), ΔE (= $f(X_i(t+1)) - f(Y_i(t))$ is calculated. Then, compare the common transition function $p(T) = exp(-\Delta E / T)$ with P_{random}. If $p(T)$ is greater than P_{random}, the next *P*best ($Y_i(t+1)$) is equal to the position of the next iteration ($X_i(t+1)$). Otherwise, the next *P*best uses the current *P*best. In this search process, the SA accepts not only better but also worse neighboring solutions with a certain probability. Such mechanism can be regarded as a trial to explore new space for new solutions, either better or worse. The probability of accepting a worse solution is larger at higher initial temperature. As the temperature decreases, the probability of accepting worse solutions gradually approaches

zero. This feature means that the SA technique makes it possible to jump out of a local optimum to search for the global optimum. The equation (4), (5), (6) and (7) below denotes the application of SA to PSO clustering.

3.2 Adjustment Algorithm

Our proposed Adjusted SA-PSO is a hybrid clustering algorithm which apply the mixture of conventional PSO and SA algorithm to the process of K-means clustering in order to solve the problem of premature convergence. Moreover we add an adjustment algorithm in the SA-PSO. This algorithm adjusts the PSO's acceleration constants (C_1 and C_2), which prevent SA-PSO from deteriorating calculation speed and improve performance. As mentioned above (Section 2.2), the fixed constants (w = 0.72 and $C_1 = C_2 = 1.49$) ensured good convergence [15]. However, we abandon the values ($C_1 = C_2 = 1.49$) and developed a new adjustment algorithm. C_1 is related to *G*best while C_2 to *G*best in the equation (8) and (9) below. After comparing *G*best with *P*best, the greater side of them has the greater related constant value by adding 0.05 on the existing constant so as to minimize the search of unnecessary areas and enhance performance. As shown in Fig. 2 this adjusted algorithm starts off with the initial value ($C_1 = C_2 = 2.0$) which is slightly greater than the fixed one from the existing literature.

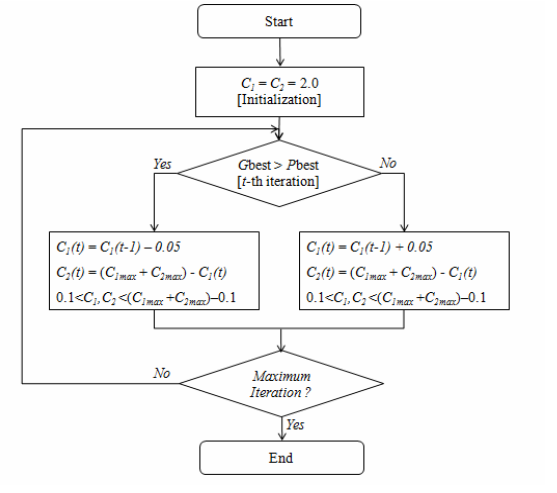


Fig. 2. Flowchart of Adjustment Algorithm

3.3 Adjusted SA-PSO Clustering

In this paper, we introduced SA into the updating of *P*best and *G*best position in PSO and adjusted algorithm into the acceleration constants in PSO. The whole procedure of Adjusted SA-PSO is described as follows:

1. K centroid vectors are chosen randomly (Initialization).
2. For t = 1 to maximum

2.1. For each particle i do (The application of K-means)

2.2. For each data vector Z_p do (The application of K-means clustering)

(Step1) Calculate the Euclidean distance $D(Z_p, C_j)$ to all G_{ij}.

The distance between each data vector and centroid vector is calculated to determine the closest centroid vector. The data vectors are clustered around each selected centroid.

$$D(Z_p, C_j) = \sqrt{\sum_{K=1}^{N_d} (Z_{p,k} - C_{j,k})^2} \tag{1}$$

K: The number of cluster centroids, i.e. the number of clusters to be formed
Z_p: The p-th data vector
C_j: The centroid vector of cluster
N_d: The input dimension, i.e. the number of parameters of each data vector

(Step2) Assign Z_p to cluster G_{ij} so as to minimize the distance D.

The average of each cluster of K subordinate data vectors is calculated and re-calculated until the criterion for centroid is satisfied.

$$C_j = \frac{1}{N_j} \sum_{\forall Z_p \in G_j} Z_p \tag{2}$$

N_j: The number of data vectors in cluster
G_j: The subset of data vectors that form cluster

(Step3) Calculate the fitness.

$$\text{Fitness function} = \frac{\sum_{j=1}^{Nc} \left(\sum_{\forall Z_p \in G_{ij}} D(Z_p, C_j) / |G_{ij}| \right)}{Nc} \tag{3}$$

N_c: The number of data vectors to be clustered

2.3. Update the global best and local best positions. (The application of SA)

(Step1) Initialize the system temperature T and the related constants.

(Step2) Repeat the following sub-steps until the criterion condition is met.

i) Update Pbest

$$\begin{cases} \text{If } f(X_i(t+1)) \le f(Y_i(t)) \text{ then } Y_i(t+1) = X_i(t+1) & (4) \\ \text{If } f(X_i(t+1)) > f(Y_i(t)) \text{ then } \Delta E = f(X_i(t+1)) - f(Y_i(t)) \\ \text{If } exp(-\Delta E / T) > P_{random}, P_{random} \in [0,1] \text{ then } Y_i(t+1) = X_i(t+1) & (5) \end{cases}$$

else $Y_i(t+1) = Y_i(t)$

X_i : The current position of the particle

Y_i : The personal best position of the particle (Pbest)

ii) Update *G*best.

$$\hat{Y}(t+1) = min\{Y_i(t+1) \mid i = 1,2,3,\dots,L\} \tag{6}$$

$\hat{Y}$: The global best position of the particle (*G*best)

iii) Cool through slowly lowering the temperature T .

$$T_{n+1} = \alpha\, T_n \tag{7}$$

α: The cooling coefficient (α< 1)

iv) Apply the adjustment algorithm.

If *G*best > *P*best (8)

then $C_1(t) = C_1(t\text{-}1) - 0.05,\ C_2(t) = (C_{1max} + C_{2max}) - C_1(t)$

If *G*best ≤ *P*best

then $C_1(t) = C_1(t\text{-}1) + 0.05,\ C_2(t) = (C_{1max} + C_{2max}) - C_1(t)$

C_1 , C_2: The acceleration constants,

(Step3) Stop if the termination condition is met.

2.4. Update the centroids by using the following equation (9) and (10).

$$V_i(t+1) = w \cdot V_i(t) + c_1 \cdot r_1(t) \cdot (Y_i(t) - X_i(t)) + c_2 \cdot r_2(t) \cdot (\hat{Y}_i(t) - X_i(t)) \tag{9}$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{10}$$

V_i : The current velocity of the particle

w : The inertia weight

$r_1(t), r_2(t) \sim U(0,1)$

2.5. Recalculate the fitness by using the above equation (1), (2) and (3).

3. Repeat until *t* becomes maximum.

4 Simulation Results and Analysis

4.1 Datasets and Setting-Up for Experiment

To measure the performance of the proposed Adjusted SA-PSO algorithm, we obtained two datasets (Bodyfat and Pollution) from the online statistics library provided by the Carnegie-Mellon University (http://lib.stat.cmu.edu/). As the two datasets has more variables than the others in that online library, their complexity might guarantee the experiment's soundness of clustering to measure the distances between particles. By using these datasets, we simulated the three algorithms that was discussed above, and analyzed the usefulness and reliability of data clustering. The bodyfat[1] dataset for

[1] As Lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men, the data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

experiment consists of 15 variables and 252 records while the pollution[2] dataset has 16 fields and 60 cases. In this experiment, we clustered the dataset into three groups (K) and set up iterations (t) 50, inertia weight (w) 0.72 and anneal constant 0.95 (α=0.95). Moreover to modify the acceleration constants, the experiment initialized 2.0 and adjusted +0.05 or -0.05 by comparing Gbest with Pbest as the number of iteration went by. Furthermore, to measure performance, 50 simulations were repeated for each algorithm (K-PSO, SA-PSO, Adjusted SA-PSO).

4.2 Usefulness and Reliability Evaluation

For the datasets of Bodyfat and Pollution, we implemented 50 simulations of the PSO, SA-PSO, and Adjusted SA-PSO algorithms and calculated the statistics for the 50-th value of fitness function in the Equation (3) above. The results are listed in the following table 1 and 2. The mean of fitness function, as our objective function, denotes the usefulness index, while MAE (Mean absolute error) represents the reliability index.

Table 1. Statistics for Fitness Function of Bodyfat Dataset

Algorithm	Mean	MAE	Variance	Max	Min	Median
K-PSO	44.087	5.118	80.771	94.693	34.955	43.069
SA-PSO	43.845	5.187	75.058	92.905	35.195	42.078
Adjusted SA-PSO	43.697	4.641	34.163	64.679	36.624	42.791

Table 2. Statistics of Fitness Function for Pollution Dataset

Algorithm	Mean	MAE	Variance	Max	Min	Median
K-PSO	1536.63	289.349	175270.15	3131.89	1107.23	1437.82
SA-PSO	1495.38	300.689	176007.61	3158.28	1063.43	1397.91
Adjusted SA-PSO	1462.09	273.997	120854.90	2677.05	1086.29	1342.24

- Usefulness Evaluation

Fitness function represents the mean of Euclidean distance. When there is little change in the centroid vectors over the number of iterations by minimizing the distance, the clustering process can he stopped. As shown in fig. 3, our new approach (Adjusted SA-PSO) had the most excellent performance (the smallest values) in usefulness evaluation.

[2] This is the pollution data so loved by writers of papers on ridge regression. Source: McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', Technometrics, vol.15, 463-482.

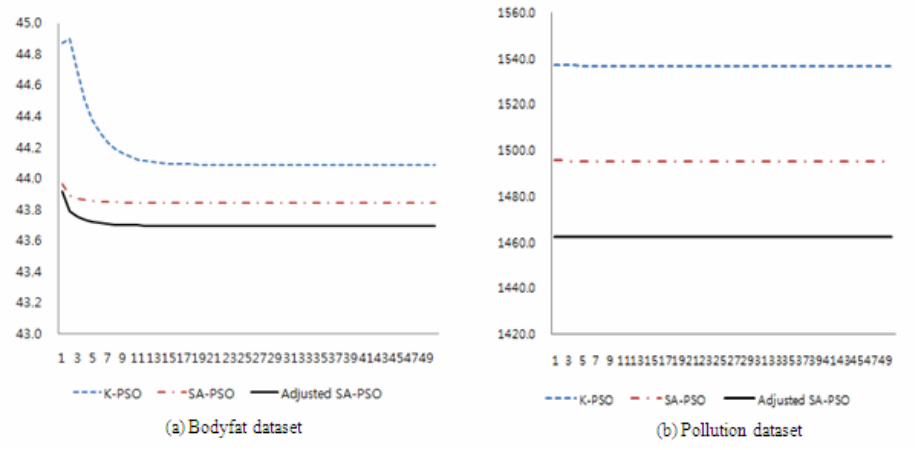


Fig. 3. Comparison of algorithms for fitness function

- Reliability Evaluation

We evaluated reliability using MAE (Mean absolute error) in the equation (11) below. The results demonstrated Adjusted SA-PSO had the best reliability (the smallest deviation) evaluation in the table 1 and 2.

(11)

F : Fitness mean
F_i: i-th fitness value
n : the number of simulations
i : i-th iteration

$$\frac{\sum_{t=1}^{n}\left|F - F_i\right|}{n}$$

5 Conclusion

This study proposed a new optimization algorithm (Adjusted SA-PSO) that has improved the conventional PSO algorithm. To test the proposed algorithm for clustering, we simulated K-PSO, SA-PSO, and ASA-PSO algorithms by using open datasets. The results showed that the ASA-PSO algorithm had the more excellent performance (minimum value) in fitness function and mean absolute error than K-PSO and SA-PSO. The Adjusted SA-PSO improved the performance of the conventional PSO by searching a wider area of solutions in the process of updating local solutions. Furthermore, it minimized the search of unnecessary areas and enhanced performance by adding the adjusted algorithm technique. For future studies, various usefulness and reliability indexes will be developed. The algorithms will also be extended to dynamically determine the optimal number of clusters.

Acknowledgments. This study was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Beyer, H.G., Schwefel, H.P.: Evolution strategies: A comprehensive introduction. Natural Computing 1(1), 3–52 (2002)
2. Da, Y., Xiurun, G.: An improved PSO-based ANN with simulated annealing technique. Neurocomputing 63, 527–533 (2005)
3. De Jong, K.A.: Are Genetic Algorithms Function Optimizers? In: Manner, R., Manderick, B. (eds.) Parallel Problem Solving from Nature 2. North-Holland, Amsterdam (1992)
4. Geman, S., Geman, D.: Stochastic Relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. PAMI 5, 721–741 (1984)
5. Huang, M.D., Romeo, F., Sangiovanni-Vincentalli, A.: An efficient general cooling schedule for simulated annealing. In: Proceedings of the IEEE International Conference on Computer Aided Design, Santa Clara, pp. 381–384 (1986)
6. Janson, S., Merkle, D., Middendorf, M.: Molecular docking with multi-objective Particle Swarm Optimization. Applied Soft Computing 8(1), 666–675 (2008)
7. Jarboui, B., Damak, N., Siarry, P., Rebai, A.: A combinatorial particle swarm optimization for solving multi-mode resource-constrained project scheduling problems. Applied Mathematics and Computation 195(1), 299–308 (2008)
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proc. IEEE Int. Conf. Neural Networks IV, pp. 1942–1948 (1995)
9. Kirkpatric, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science 220, 671–680 (1983)
10. van Laarhoven, P.J.M., Aarts, E.H.L.: Simulated Annealing: Theory and Applications (1988)
11. Liang, J.J., Qin, A.K., Suganthan, P.N., Baskar, S.: Comprehensive learning particle swarm optimizer for global 11. optimization of multimodal functions. IEEE Transactions on Evolutionary Computation 10(3), 281–295 (2006)
12. Liu, B., Wang, L., Jin, Y.: An effective hybrid PSO-based algorithm for flow shop scheduling with limited buffers. Computers & Operations Research 35(9), 2791–2806 (2008)
13. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium and Mathematical Statistics and Probability, vol. 1, pp. 281–296 (1967)
14. Maitra, M., Chatterjee, A.: A hybrid cooperative–comprehensive learning based PSO algorithm for image segmentation using multilevel thresholding. Expert Systems with Applications 34(2), 1341–1350 (2008)
15. Van den Bergh, F.: An analysis of particle swarm optimizers. PhD Thesis, Department of Computer Science, University of Pretoria, Pretoria, South Africa (2002)
16. Van der Merwe, D.W., Engelbrecht, A.P.: Data clustering using particle swarm optimization. In: The 2003 Congress on Evolutionary Computation, pp. 215–220 (2003)
17. Wang, D., Liu, L.: Hybrid particle swarm optimization for solving resource-constrained FMS. Progress in Natural Science 18(9), 1179–1183 (2008)

Author Index